Escuela Politécnica Superior

Departamento de Teoría de la Señal y Comunicaciones

PROYECTO FIN DE CARRERA

APRENDIZAJE MÁQUINA PARA LA AYUDA DE TOMA DE DECISIONES EN LA GESTIÓN DE UNA UNIDAD DE CUIDADOS INTENSIVOS

Autor: DAVID TOLEDO NAVARRO

Tutor: EMILIO PARRADO HERNÁNDEZ

Leganés, marzo de 2013

Título: APRENDIZAJE MÁQUINA PARA LA AYUDA DE TOMA DE DECISIONES EN LA GESTIÓN DE UNA UNIDAD DE CUIDADOS INTENSIVOS

Autor: DAVID TOLEDO NAVARRO
Director: EMILIO PARRADO HERNÁNDEZ

EL TRIBUNAL

Presidente: ANGEL NAVIA VÁZQUEZ

Vocal: FERNANDO FERNÁNDEZ REBOLLO

Secretario: RUBÉN SOLERA UREÑA

Realizado el acto de defensa y lectura del Proyecto Fin de Carrera el día 21 de marzo de 2013 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de

VOCAL

SECRETARIO PRESIDENTE

Agradecimientos

Quiero agradecer en primer lugar a Emilio Parrado, tutor de este proyecto, su dedicación y toda la ayuda prestada.

A todos mis compañeros de la UC3M, por los buenos momentos compartidos.

A mis amigos *del pueblo*, que tantas veces he echado de menos, por su apoyo y ánimo desde la distancia.

A ti Elisa, por la compresión y paciencia que has tenido conmigo, y todo el cariño que he recibido de tu parte. Te quiero mucho, Elisa.

Todas las palabras son pocas para agradecer el apoyo incondicional y el cariño de mi familia. A mis padres Fileto y Esperanza, y mis hermanos Esperanza, Julia, Sebastián, Ana María y Raquel, a los que nunca podré agradecer lo suficiente todas las oportunidades que me han dado y su apoyo constante durante todos estos años. La alegría de mis sobrinos Leonardo, Gonzalo y Abel, y la de todos los *pequeñajos* que están por venir.

A mis padres y mis hermanos,
con todo mi cariño

Índice general

Índice de figuras

Índice de tablas

Resumen

En este proyecto se realiza un estudio experimental relativo a la aplicación de la máquina de vectores soporte (*Support Vector Machine*, SVM) para la predicción de la mortalidad de los pacientes en el servicio de la Unidad de Cuidados Intensivos (UCI), con el objetivo de estudiar si existe o no una clase definida para los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad, dado por el índice APACHE (*Acute Physiology And Chronic Health Evaluation*) II, bajo, determinando si los resultados obtenidos del estudio son significativos en el sentido de ayuda a la toma de decisiones respecto a los pacientes que mueren en el servicio de la UCI.

En todas las terapias intensivas se utiliza este índice como marcador pronóstico al ingreso de los pacientes críticos, lo que permite estratificar la complejidad de los pacientes internados, observando, en base a los datos obtenidos, que existe una relación directamente proporcional entre el índice APACHE II y la mortalidad. Sin embargo, las circunstancias particulares de cada hospital (características demográficas de los pacientes, capacitación del personal, etc.) hacen que las predicciones efectuadas por el índice APACHE II no siempre se cumplan. Por otra parte, cada vez son más las patologías donde este índice es un marcador independiente de la mortalidad, como es en el caso de las pancreatitis.

La máquina de vectores soporte, en su modalidad para clasificación (*Support Vector Classifier*, SVC), se utiliza para predecir a posteriori la mortalidad de cada uno de los pacientes de la UCI, es decir, saber si un paciente tiene una evolución favorable en el servicio de la UCI o, por el contrario, muere, comparando los resultados obtenidos mediante un procedimiento de mínimos cuadrados (*Least Squares*, LS) SVC y la descripción de datos por vectores soporte (*Support Vector Data Description*, SVDD), respecto al resultado que se puede obtener utilizando el índice APACHE II como estimador *baseline* (sistema de referencia). La máquina de vectores soporte tiene un modelo estado del arte no superado hasta el momento, SVC, que le confieren a priori ciertas ventajas respecto a otras técnicas empleadas, obteniendo, una vez que se han fijado adecuadamente los parámetros de la SVM, excelentes precisiones y buenas propiedades de generalización. Con el sistema LS-SVC propuesto se consiguen resultados competitivos respecto al resultado obtenido con el sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II.

Finalmente, se proponen una serie de líneas futuras de investigación que son de especial interés.

Palabras clave: UCI, APACHE II, SVM, LS-SVC, SVDD, Matlab.

Abstract

This project is a pilot study on the application of Support Vector Machine (SVM) for prediction of mortality in patients in the service of the Intensive Care Unit (ICU), with the objective to study whether there is a class defined for patients dying in the ICU have a mortality predictor, given by the APACHE (Acute Physiology and Chronic Health Evaluation) II, low, determining whether the results of the study are significant in that helps decision making regarding patients dying in the ICU service.

In all intensive therapies used this index as a prognostic marker admission of critically ill patients, allowing stratify the complexity of the patient population, observing, based on the data obtained, there is a direct relationship between the APACHE II and mortality. However, the particular circumstances of each hospital (demographic characteristics of patients, staff training, etc.) make the predictions made by the APACHE II is not always met. Moreover, an increasing pathology, this index is an independent predictor of mortality, as in the case of pancreatitis.

The support vector machine, as it applies to classification (Support Vector Classifier, SVC), is used to predict subsequent mortality in each ICU patients, whether a patient has a favorable trend in the ICU service or, conversely, dies, comparing results obtained

using a Least Squares (LS) SVC method and the Support Vector Data Description (SVDD), regards the result to can be obtained using the APACHE II and baseline estimator (reference system). The support vector machine is a model state of the art so far not exceeded, SVC, giving it a priori certain advantages over other techniques used, obtaining, once they have set the parameters appropriately SVM, excellent precision and good generalization properties. With the proposed LS-SVC achieve competitive results compared to the results obtained with the reference system based on soft output used as the classifier the APACHE II.

Finally, we outline some future lines of research that are of special interest.

Keywords: ICU, APACHE II, SVM, LS-SVC, SVDD, Matlab.

Siglas

APACHE	*Acute Physiology And Chronic Health Evaluation*
CMV	*Continuous Mandatory Ventilation*
GCS	*Glasgow Coma Score*
IDE	*Integrated Development Environment*
IVA	Impuesto sobre el Valor Agregado
KKT	Karush-Kuhn-Tucker
LS	*Least Squares*
OHDR	*Optimal Hyperplane Decision Rule*
OR	*Odds Ratio*
PDI	Personal Docente e Investigador
PEEP	*Positive End Expiratory Pressure*
QP	*Quadratic Programming*
ROC	*Receiver Operating Characteristic*
SIDA	Síndrome de Inmunodeficiencia Adquirida
STD	*Standard Deviation*
SVC	*Support Vector Classifier*
SVM	*Support Vector Machine*
TAC	Tomagrafía Axial Computarizada

UCI	Unidad de Cuidados Intensivos

Capítulo 1

1. Introducción y objetivos

En este capítulo se describe, de forma general, este Proyecto Fin de Carrera, prestando especial atención a la motivación del trabajo realizado, y los objetivos planteados del mismo, así como a la presentación de las principales contribuciones técnicas realizadas. También, en este capítulo, se hace mención de las diferentes fases en las que se ha desarrollado el proyecto, los medios que se han empleado, y la descripción de los capítulos que forman la estructura de la memoria.

1.1. Motivación del proyecto

La mortalidad en los pacientes admitidos en el servicio de la Unidad de Cuidados Intensivos (UCI) de un hospital es mucho más alta que la de otros pacientes hospitalarios. Considerando la mortalidad relativamente alta entre estos pacientes, la mortalidad es entonces una medida sensible, apropiada, y útil para medir resultados, que dependen no solo del personal médico, de los equipos utilizados, y de los procesos de

cuidado, sino también de las características de los pacientes admitidos. La UCI admite mayores proporciones de pacientes de alto riesgo, por lo cual, debería esperarse una mayor mortalidad [GR99]. La falta de exactitud en el modelo de predicción de mortalidad, por sí mismo, puede ser responsable de la variación entre la mortalidad observada y predicha [LFT00].

En este proyecto, con la autorización y asesoría de los coordinadores de la UCI del Hospital Clínico Universitario de Valladolid, se realiza un amplio estudio sobre un histórico de datos de 1100 pacientes, debidamente procesado para preservar el anonimato de dichos pacientes. En esta base de datos aparecen variables relacionadas con la evolución de los pacientes en la UCI, así como el índice APACHE (*Acute Physiology And Chronic Health Evaluation*) II [KDW+85], que se le asignó en el ingreso. El índice APACHE II es uno de los sistemas más frecuentemente utilizados para cuantificar la gravedad de un paciente con independencia del diagnóstico. En base a este índice se puede predecir la evolución de los pacientes por medio de una cifra objetiva. La ecuación que proporciona este índice se ha obtenido mediante la compilación de conocimiento experto y el estudio estadístico de 18000 casos en hospitales de EE. UU. Sin embargo, las circunstancias particulares de cada hospital (características demográficas de los pacientes, capacitación del personal, etc.) hacen que las predicciones efectuadas por el índice APACHE II no siempre se cumplan. Son de interés los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo. Por lo tanto, la motivación de este proyecto es predecir la mortalidad de cada uno de los pacientes de la UCI mediante técnicas basadas en la máquina de vectores soporte.

Cada una de las filas de la base de datos proporcionada por el Hospital Clínico Universitario de Valladolid corresponde a un paciente, y en las columnas se sitúan las variables que caracterizan a cada uno de los pacientes de la UCI, como son: la fecha de nacimiento, sexo, fecha de ingreso en el hospital, fecha de ingreso en la UCI, procedencia, causa de ingreso en el hospital, enfermedad de ingreso en la UCI, índice APACHE II, antecedentes, diagnósticos, técnicas utilizadas, complicaciones, fecha de alta del servicio de la UCI, y motivo del alta. Variables que se describen con detalle en el

Capítulo 4. Un problema común y fundamental sobre el que construir los modelos de los escenarios citados anteriormente es la adopción de una medida de disimilitud o similitud entre pacientes, representados por las variables de la base de datos. Estas variables son tanto binarias como cualitativas y cuantitativas, por lo tanto, es necesario el estudio de índices de similitud para un conjunto de variables mixtas.

La máquina de vectores soporte, en su modalidad para clasificación (*Support Vector Classifier*, SVC), se utiliza para predecir a posteriori la mortalidad de cada uno de los pacientes de la UCI, es decir, saber si un paciente tiene una evolución favorable en el servicio de la UCI o, por el contrario, muere, comparando los resultados obtenidos mediante LS-SVC y SVDD, respecto al resultado que se puede obtener utilizando el índice APACHE II como estimador *baseline*.

La máquina de vectores soporte (*Support Vector Machine*, SVM) [BGV92] [Vap95] es una herramienta que ha demostrado en los últimos años excelentes resultados en una gran variedad de problemas de clasificación. Junto a esto, su sólida base teórica ha llevado a considerar a las SVMs como los actuales modelos estado del arte para clasificación, ya que las siguientes características de la máquina de vectores soporte le confieren a priori ciertas ventajas respecto a otras técnicas empleadas:

- Máximo compromiso entre generalización y precisión (controlado por un factor de regularización).
- Al contrario de lo que ocurre con otras aproximaciones no lineales, en la SVM está garantizada la existencia y unicidad de la solución óptima. Esto se debe a que el funcional que se minimiza es siempre una forma cuadrática.
- La máquina de vectores soporte puede tratar con muestras de entrada de muy alta dimensión, en virtud del clásico *truco del kernel* [ABR64].
- Se obtienen modelos resultantes dispersos, es decir, tras el entrenamiento se descarta parte de las muestras de entrada.

No obstante, también hay que señalar que la máquina de vectores soporte presenta una serie de inconvenientes como los que se destacan a continuación:

- El crecimiento coste computacional de aprendizaje de la SVM, en términos de tiempo y memoria, con el número de muestras de entrenamiento, impidiendo el uso de grandes bases de datos.
- La optimización de los parámetros no lineales de la SVM (usualmente, factor de regularización y parámetros de la función de *kernel*) hace que el entrenamiento sea mucho más complejo y costoso, implicando un problema de optimización global (en [SVD02], se ve cómo se puede abordar por niveles la optimización, pero sigue siendo muy costoso).

En resumen, para clasificación la máquina de vectores soporte tiene un modelo estado del arte no superado hasta el momento, SVC, obteniendo, una vez que se han fijado adecuadamente los parámetros de la SVM, excelentes precisiones y buenas propiedades de generalización.

1.2. Objetivos

1.2.1. Objetivo principal

Una vez presentados el contexto y la motivación del proyecto, cabe señalar que, **el objetivo final de este proyecto es estudiar si existe o no una clase definida para los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo,** determinando si los resultados obtenidos del estudio son significativos en el sentido de ayuda a la toma de decisiones respecto a los pacientes que mueren en el servicio de la UCI del Hospital Clínico Universitario de Valladolid, y si sería posible generalizar dichos resultados a cualquier UCI de otros hospitales.

1.2.2. Objetivos secundarios

En base a este objetivo principal se propone como objetivo parcial la obtención de diferentes indicadores, como resultado del estudio estadístico de las variables que caracterizan a cada uno de los pacientes, que puedan ser de interés para su utilización en la UCI de dicho hospital.

1.3. Fases del desarrollo

Se realiza un estudio prospectivo, utilizando diferentes esquemas de aprendizaje máquina y minería de datos, donde se adopta una metodología con los siguientes pasos: (1) descripción y planteamiento del problema a resolver, (2) importación y codificación de los datos, (3) selección del conjunto de datos, (4) análisis de las propiedades de los datos, (5) transformación del conjunto de datos de entrada, (6) propuesta y elección de las técnicas y algoritmos a implementar, (7) estudio de reconocimiento supervisado mediante clasificación, utilizando clasificadores SVM, y (8) evaluación de los resultados obtenidos y conclusiones.

A continuación se explican cada una de las fases en las que se ha desarrollado el proyecto, enumeradas en la presentación del proyecto.

1. **Descripción y planteamiento del problema a resolver**. Los investigadores del Hospital Clínico Universitario de Valladolid presentaron una base de datos, en Excel, de un histórico de medidas biomédicas recopiladas de 1100 pacientes de la UCI de dicho hospital. A partir de dicha información proporcionada y las especificaciones dadas, se describen una serie de posibles líneas de investigación a estudiar.
2. **Importación y codificación de los datos**. La hoja Excel se importa a Matlab para posteriormente trabajar mejor la parte de algoritmia. Existió un compromiso de

colaboración y apoyo de los investigadores del Hospital Clínico Universitario de Valladolid para entender e interpretar los diferentes registros de la base de datos. Es entonces cuando se realiza un proceso típico de minería de datos [Che10].

3. **Selección del conjunto de datos**. Tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.
4. **Análisis de las propiedades de los datos**. En especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
5. **Transformación del conjunto de datos de entrada**. Se realiza la codificación de los datos con un formato específico, como se explica en detalle en el Capítulo 4 de la memoria del proyecto, con el objetivo de prepararlos para aplicar las técnicas que mejor se adapten a los datos y al problema. A este paso también se le conoce como pre procesamiento de los datos.
6. **Propuesta y elección de las técnicas y algoritmos a implementar.** Se hace una valoración de las diferentes técnicas de aprendizaje máquina y algoritmos de minería de datos, ya que la naturaleza de los datos juega un papel muy importante en la precisión de la clasificación, y es necesaria su elección para un conjunto de datos en particular [Mar08]. Esta valoración concluye con la propuesta de un estudio de reconocimiento supervisado mediante técnicas de clasificación. Un problema común y fundamental sobre el que construir los modelos de los escenarios citados anteriormente es la adopción de una medida de disimilitud o similitud entre pacientes, representados por las variables de la base de datos. Estas variables son tanto binarias como cualitativas y cuantitativas, por lo tanto, es necesario el estudio de índices de similitud para un conjunto de variables mixtas.
7. **Estudio de reconocimiento supervisado**. En este caso se quiere predecir la mortalidad, desconocida a priori, a partir de las variables que caracterizan a cada uno de los pacientes. Haciendo uso de la librería de Matlab *bioinformatics toolbox*, se utilizan técnicas de clasificación, mediante el uso de máquinas de vectores soporte (SVM), comparando el resultado de precisión promedio de clasificación obtenido mediante LS-SVC y SVDD, respecto a la precisión

promedio que se puede obtener si se utiliza como salida blanda del clasificador el índice APACHE II (estimador *baseline*). El indicador de precisión promedio de clasificación, como se define en el Capítulo 4 de la memoria del proyecto, es una medida de calidad del clasificador SVM. Son de interés los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo.

8. **Evaluación de los resultados obtenidos y conclusiones**. Una vez se han obtenido los diferentes resultados de los modelos implementados, se realiza una profunda evaluación de los mismos para deducir las conclusiones finales.

1.4. Medios empleados

Para la elaboración de este proyecto se ha utilizado el paquete ofimático Microsoft Office 2011, como se ha comentado en las secciones anteriores, se ha trabajado con una base de datos en Excel, y los diferentes modelos de aprendizaje máquina y minería de datos se han implementado en el lenguaje de programación Matlab.

MATLAB® (abreviatura de *MATrix LABoratory*, "laboratorio de matrices") es un software matemático que ofrece un entorno de desarrollo integrado (IDE) con un lenguaje de programación propio (lenguaje M) de alto nivel para el cálculo numérico, la visualización y programación, permitiendo realizar tareas computacionales intensivas [MATHWORKS].

1.5. Estructura de la memoria

Para facilitar la lectura de la memoria, se incluye, a continuación, un breve resumen de cada capítulo. La estructura de la memoria es la siguiente: en el Capítulo 2 se realiza un estudio del índice APACHE II en terapia intensiva donde se obtiene la correlación del índice APACHE II con la mortalidad al ingreso de los pacientes de la UCI, y se establece

el punto de corte del índice APACHE II que se comporte como marcador independiente de mortalidad. En el Capítulo 3 se presentan los fundamentos de la máquina de vectores soporte, prestando especial atención a los aspectos más relevantes para el trabajo desarrollado posteriormente. A continuación, en el Capítulo 4 y en el Capítulo 5 se describen el trabajo realizado para obtener la predicción de la mortalidad de los pacientes de la UCI mediante SVCs, incluyendo una discusión de los resultados experimentales obtenidos. En el Capítulo 6 se revisan las contribuciones técnicas propuestas en este proyecto, se exponen las conclusiones del estudio realizado y se discuten las posibles líneas futuras de investigación. Por último, en el Capítulo 7 se muestra la planificación que se ha seguido para el desarrollo del proyecto y la valoración económica resultante.

Capítulo 2

2. Índice APACHE II

En este capítulo se presenta y describe el índice APACHE (*Acute Physiology And Chronic Health Evaluation*) II [KDW+85] utilizado en terapia intensiva. A continuación, se realiza un estudio del índice APACHE II medido en los pacientes de la UCI de la base de datos aportada por el Hospital Clínico Universitario de Valladolid, donde se obtiene la correlación del índice APACHE II con la mortalidad al ingreso de los pacientes de la UCI, y se establece el punto de corte del índice APACHE II que se comporte como marcador independiente de mortalidad.

2.1. Introducción

Como se ha presentado en la introducción del proyecto, el índice APACHE II es un sistema de valoración pronostica de mortalidad, que consiste en detectar los trastornos fisiológicos agudos que atentan contra la vida del paciente, y se fundamenta en la determinación de las alteraciones de variables fisiológicas y de parámetros de laboratorio,

cuya puntuación es un factor predictivo de mortalidad, siendo este índice válido para un amplio rango de diagnósticos, fácil de usar y que puede sustentarse en datos disponibles en la mayor parte de las UCI.

Variables fisiológicas	Rango elevado				Rango bajo				
	+4	+3	+2	+1	0	+1	+2	+3	+4
Temperatura rectal (axial +0.5°C)	≥ 41	39–40.9°		38,5-38,9°	36-35,9°	34-35,9°	32-33,9°	30-31,9°	≤29,9°
Presión arterial media (mmHg)	≥ 160	130-159	110-129		70-109		50-69		≤49
Frecuencia Cardiaca (respuesta ventricular)	≥ 180	140-179	110-139		70-109		55-69	40-54	≤39
Frecuencia respiratoria (no ventilado o ventilado)	≥ 50	35–49		25-34	12-24	10-11	6-9		≤5
Oxigenación: elegir a o b									
a. si FiO2 ≥0,5 anotar PA-aO2	≥ 500	350-499	200-349		<200				
b. si FiO2 < 0,5 anotar PaO2					> 70	61-70		55-60	≤55
*Ph arterial (preferido)	≥ 7.7	7.6–7.59		7,5-7,49	7,33-7,49		7,25-7,32	7,15-7,24	<7.15
*HCO3 sérico (venoso mEq/l)	≥ 52	41-51.9		32-40,9	22-31,9		18-21,9	15-17,9	<15
Na+ sérico (mEq/l)	≥ 180	160-179	155-159	150-154	130-149		120-129	111-119	≤110
K+ sérico (mEq/l)	≥ 7	6–6.9		5,5-5,9	3,5-5,4	3-3,4	2,5-2,9		<2,5
*Creatinina sérica (md/dl)	≥ 3.5	2-3,4	1,5-1,9		0,6-1,4		<0,6		
*Doble puntuación en caso de fallo renal agudo									
Hematocrito (%)	≥ 60		50-59,9	46-49,9	30-45,9		20-29,9		<20
Leucocitos (total/mm3 en miles)	≥ 40		20-39,9	15-19,9	3-14,9		1-2,9		<1
Escala de Glasgow									
Puntuación=15- Glasgow actual									
A. APS (Acute Physiology Score) Total: suma de las 12 variables individuales									
B. Puntuación por edad (≤ 44 = 0 punto; 45-54 = 2 puntos; 55-64 = 3 puntos; 65-74 = puntos; >75 = 6 puntos)									
C. Puntuación por enfermedad crónica									
Puntuación APACHE II (suma de A+B+C)									

Puntuación por enfermedad crónica: Si el paciente tiene historia de insuficiencia orgánica sistémica o está inmunocomprometido, corresponde 5 puntos en caso de postquirúrgicos urgentes o no quirúrgicos, y 2 puntos en caso de postquirúrgicos de cirugía electiva [KDW+85].

Definiciones: Debe existir evidencia de insuficiencia orgánica o inmunocompromiso, previa al ingreso hospitalario y conforme a los siguientes criterios:

- Hígado: Cirrosis (con biopsia), hipertensión portal comprobada, antecedentes de hemorragia gastrointestinal alta debida a Hipertensión Arterial (H. T. A.) portal o episodios previos de fallo hepático, encefalohepatopatía, o coma.
- Cardiovascular: Clase IV según la New York Heart Association.
- Respiratorio: Enfermedad restrictiva, obstructiva o vascular que obligue a restringir el ejercicio, como por ej. Incapacidad para subir escaleras o realizar tareas domésticas; o hipoxia crónica probada, hipercapnia, policitemia secundaria, hipertensión pulmonar severa (>40 mmHg), o dependencia respiratoria.
- Renal: Hemodializados.
- Inmunocomprometidos: que el paciente haya recibido terapia que suprima la resistencia a la infección (inmunosupresión, quimioterapia, radiación, tratamiento crónico o altas dosis recientes de esteroides, o que padezca una enfermedad suficientemente avanzada para inmunodeprimir como por ejemplo leucemia, linfoma, SIDA, etc.).

El índice APACHE II es calculado en el momento de ingreso o al final del día de internación del paciente, por lo tanto la misma, brinda un perfil momentáneo del estado del internado, no pudiendo aportar información dinámica [CGM99].

En todas las terapias intensivas se utiliza este índice como marcador pronóstico al ingreso de los pacientes críticos, lo que permite estratificar la complejidad de los pacientes internados. Por otra parte, cada vez son más las patologías donde este índice es un marcador independiente de la mortalidad, como es en el caso de las pancreatitis.

2.2. Estudio del índice APACHE II en terapia intensiva

Mediante un estudio retrospectivo observacional realizado sobre los pacientes de la UCI del Hospital Clínico Universitario de Valladolid, se pretende correlacionar el índice

Capítulo 2. Índice APACHE II

APACHE II con la mortalidad al ingreso de los pacientes críticos y establecer el punto de corte del índice APACHE II que se comporte como marcador independiente de mortalidad. Se excluyen a los pacientes menores de 18 años y aquellos cuya internación en la UCI duró menos de 24 horas.

Los pacientes fueron divididos en grupos, en base al valor del índice APACHE II, y se procedió a la evaluación de la mortalidad en cada uno de ellos. En la *Tabla 1* se observa la mortalidad de los grupos APACHE II.

Grupos	Puntuación APACHE II	Mortalidad (%)
1	0-4	0
2	5-9	2.24
3	10-14	8.72
4	15-19	25.58
5	20-24	45.65
6	25-29	50
7	30-34	66.67
8	35-39	71.43
9	40-44	100
10	45-49	100

Tabla 1. Relación de mortalidad según el índice APACHE II

Los valores de mortalidad según el índice APACHE II se pueden aproximar por una ecuación de probabilidad aproximada, obtenida mediante regresión logística. Su fórmula más básica (para un solo predictor) es la siguiente [HL89]:

$$\pi(x) = \frac{e^{(\beta_0+\beta_1 x)}}{e^{(\beta_0+\beta_1 x)}+1} = \frac{1}{e^{-(\beta_0+\beta_1 x)}+1} \tag{2.1}$$

siendo, en este caso, que $\beta_0 = -3.85$ y $\beta_1 = 0.76$. En la *Figura 1* se observa la función de probabilidad aproximada de mortalidad obtenida según el índice APACHE II.

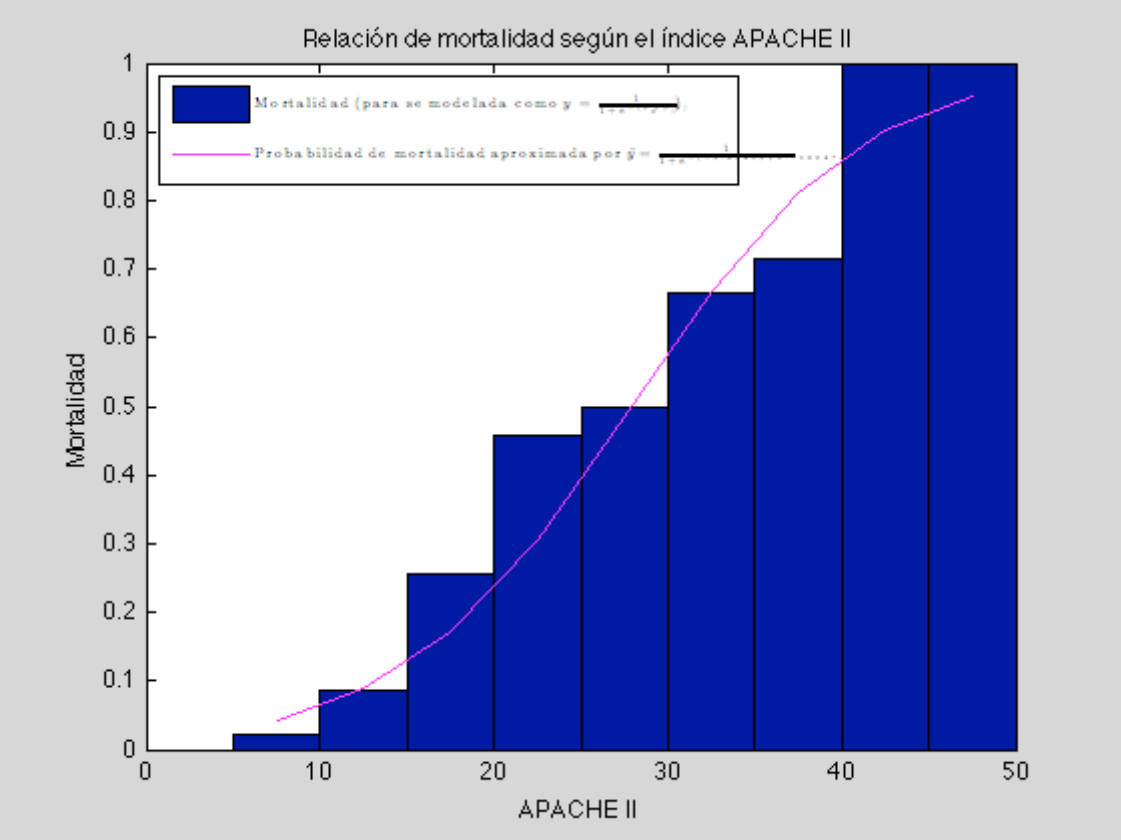

Figura 1. Relación de mortalidad según el índice APACHE II

Costa, J. I. Y Gomes do Amaral, J. L. en [CLJ89] estudiaron la relación entre el índice APACHE II y la evolución de los pacientes internados en la UCI, observando que a menor valor de la escala, mayor es la sobrevida de los pacientes. En este caso, se puede afirmar, en base a los datos obtenidos, que existe una relación directamente proporcional entre el índice APACHE II y la mortalidad.

Por otro lado, el punto de corte del índice APACHE II que predice la mortalidad, según la curva ROC, es de 26. En la *Figura 2* se muestra la representación de la curva ROC para el índice APACHE II.

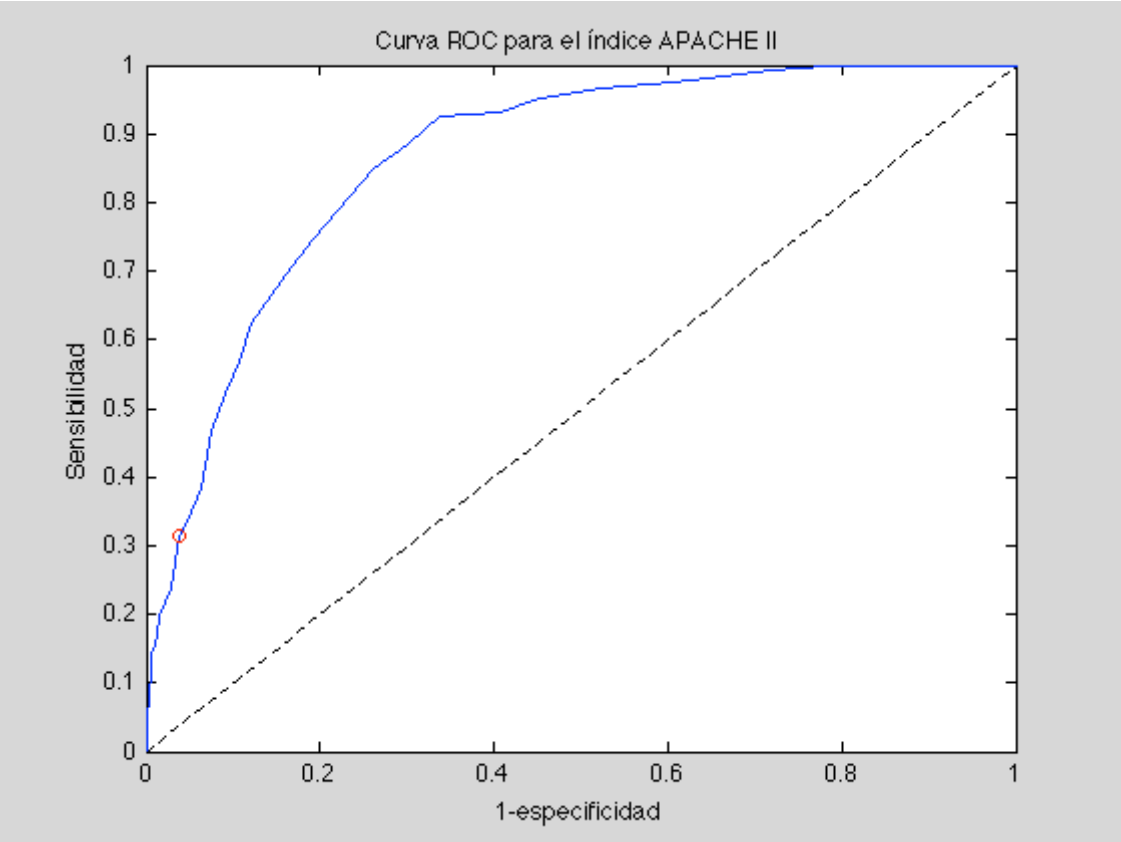

Figura 2. Curva ROC para el índice APACHE II

En el análisis, el *Odds Ratio* (OR) de mortalidad es de 11.64 con un intervalo de confianza del 95% de: (6.88, 19.67), lo que significa que aquellos pacientes con índice APACHE II mayor de 26 tendrán más probabilidades de evolucionar hacia el óbito.

En el estudio de Rioseco, M. L. y Riquelme, R. O. [RR04], donde se evaluaron a pacientes con neumonía neumocócica grave, el valor de un índice APACHE II mayor de 16 se relacionó con una mayor mortalidad. Lesage y Ramakers en [LRD04], realizaron un estudio acerca de los factores pronósticos en pacientes con infarto agudo de

miocardio, obteniendo que los pacientes con valores de índice APACHE II mayor de 29 presentaron una mayor mortalidad. Si bien, la literatura publicada demuestra que los puntos de corte se encuentran en distintos valores de índice APACHE II al observado en este caso, debido a que responde exclusivamente a los datos de este proyecto.

Son de interés los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo.

Capítulo 3

3. Máquinas de vectores soporte

En este capítulo se presenta la máquina de vectores soporte en su modalidad de clasificación (*Support Vector Classifier*, SVC) como método para su uso en el estudio de la predicción de la mortalidad de los pacientes en el servicio de la UCI. En primer lugar, se hace una introducción de los conceptos esenciales y las herramientas matemáticas que constituyen la base teórica de la SVM. A continuación se desarrolla su formulación para la modalidad de clasificación, describiendo los modelos siguientes:

1- El método de mínimos cuadrados (*Least Squares*, LS) SVC, que son una modificación de la formalización básica de las SVMs, en las que la optimización lleva a resolver un sistema de ecuaciones lineales, más sencillo de utilizar que las soluciones a la programación cuadrática.
2- La descripción de datos por vectores soporte (*Support Vector Data Description*, SVDD), también llamada *one-class* SVC, que es un tipo especial de problema de clasificación de dos clases, donde se utilizan medidas de distancia sobre hiperesferas para delimitar los datos.

3.1. Introducción

La máquina de vectores soporte (*Support Vector Machine*, SVM) es un método de aprendizaje basado en muestras. Este algoritmo fue propuesto por Vapnik [Vap82][Vap95] para la resolución de problemas de clasificación linealmente separables mediante lo que se denomina hiperplano óptimo de separación (*Optimal Hyperplane Decision Rule*, OHDR). La máquina de vectores soporte propuesta originalmente por Vapnik y sus colaboradores, amplió su ámbito de trabajo a la resolución de problemas de clasificación no separables mediante algoritmos no lineales [BGV92][GBV93][CV95]. Posteriormente, dicha metodología se extendió para su uso en problemas de regresión [Vap95][Smo96][DBK+97].

La formulación de la SVM parte del concepto clásico de hiperplano óptimo de separación que define un clasificador binario que separa las muestras de entrenamiento de cada clase con una capacidad de generalización superior a la de otros métodos de aprendizaje. Esta mayor capacidad de generalización es consecuencia principalmente de la maximización del margen del hiperplano de separación, obtenida con el hiperplano que quede a mayor distancia de los conjuntos de muestras de entrenamiento que separan. El resultado es un método de aprendizaje que ha proporcionado excelentes resultados en una gran diversidad de problemas prácticos.

A continuación se exponen los conceptos teóricos y las herramientas matemáticas necesarios para el desarrollo de la formulación de la máquina de vectores soporte.

3.1.1. Fundamentos de la máquina de vectores soporte

Considérese un conjunto de entrenamiento *d*-dimensional linealmente separable, compuesto por n muestras $\mathbf{x}_i \in R^d \left(i = 1,...,n\right)$ y sus correspondientes etiquetas

$y_i \in \{+1, -1\}$. Al tratarse de un problema separable, existe algún hiperplano definido por su vector director $\mathbf{w}$ y el sesgo b para el que la salida blanda del clasificador cumple:

$$\left| f(\mathbf{x}_i) \right| = \left| \mathbf{w}^T \mathbf{x}_i + b \right| \geq 1; \ \forall i = 1, ..., n \tag{3.1}$$

De la expresión anterior, se obtiene que la distancia existente entre la frontera de decisión y cualquier muestra $\mathbf{x}$ es:

$$r_x = \frac{\left| f(\mathbf{x}) \right|}{\left\| \mathbf{w} \right\|} \geq \frac{1}{\left\| \mathbf{w} \right\|} \tag{3.2}$$

siendo el margen del hiperplano de separación la distancia $2/\left\| \mathbf{w} \right\|$.

El hiperplano óptimo de separación es aquél capaz de separar correctamente las muestras de entrenamiento de cada clase con el mayor margen posible. La búsqueda del OHDR se puede plantear como el siguiente problema de optimización:

$$\min_{\mathbf{w}, b} \frac{1}{2} \left\| \mathbf{w} \right\|^2 \tag{3.3}$$

$$\text{sujeto a } y_i \left(\mathbf{w}^T \mathbf{x}_i + b \right) \geq 1; \ \forall i = 1, ..., n \tag{3.4}$$

Suponiendo que $\left\| \mathbf{w} \right\| \leq A, \ A \in R^+$, el margen del clasificador debe ser mayor que $2/A$, reduciendo el número de hiperplanos de separación factibles. Esta interpretación geométrica no es tan intuitiva en regresión, considerándose en este caso que la minimización de la norma de $\mathbf{w}$ conduce a soluciones suaves y, por tanto, más robustas en condiciones ruidosas.

La maximización del margen no implica necesariamente la minimización del número de errores en problemas de clasificación no separables. En la siguiente sección se verá que el

problema de optimización (3.3) debe ser modificado para tener en cuenta el error cometido en el conjunto de entrenamiento.

3.1.2. Extensión no lineal de la máquina de vectores soporte

El interés por usar fronteras de decisión no lineales surge por las limitaciones que impone el uso de hiperplanos en problemas de clasificación o de regresión cuya solución óptima tiene forma no lineal.

La extensión no lineal de la SVM se basa en una transformación implícita de las muestras de entrada a un espacio de características (también llamado espacio de Hilbert) H de mayor dimensión (posiblemente infinita) mediante una función $\phi: R^d \rightarrow H$, que puede conducir a una mayor separación entre las distintas clases (no asegurando que las clases sean completamente separables en el espacio transformado), y sobre el que se entrena una máquina lineal para definir el hiperplano óptimo de separación. Como se verá en la siguiente sección, la formulación dual de la SVM no lineal queda expresada únicamente en función de los productos escalares de las muestras transformadas: $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$. En este caso, el *truco del kernel* [ABR64] permite eludir la necesidad de calcular de manera explícita las muestras transformadas en el espacio de características si se puede definir una función de *kernel* $K: R^d \times R^d \rightarrow R$ que represente un producto escalar en el espacio transformado: $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$ que cumpla las condiciones dadas por el teorema de Mercer [Mer09]. Dicho teorema establece que existe una transformación $\phi: R^d \rightarrow H$ y una función de *kernel* que representa un producto escalar en el espacio de características asociado $K(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j)$ si y solo si se cumple que:

$$\int_X \int_X K(\mathbf{x},\mathbf{z})\, g(\mathbf{x})\, g(\mathbf{z})\, d\mathbf{x}\, d\mathbf{z} \geq 0 \tag{3.5}$$

para cualquier función $g(\cdot)$ tal que:

$$\int_X g^2(\mathbf{x})\, d\mathbf{x} < \infty \tag{3.6}$$

Dado un conjunto finito cualquiera de muestras de entrenamiento $\left\{(\mathbf{x}_i, y_i)_{i=1}^n\right\} \in \{X \times Y\}$ la condición (3.5) se simplifica como:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_i d_j K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{d}^T \mathbf{K} \mathbf{d} \geq 0;\ \forall \mathbf{d} = [d_1, d_2, ..., d_n]^T \in R^n \tag{3.7}$$

donde $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)\ \forall i, j = 1, ..., n$. Por lo tanto, la matriz de *kernels* $\mathbf{K}$ debe ser semidefinida positiva.

La selección de una función de *kernel* adecuada depende en gran medida de las características del problema de aprendizaje que se aborda. En [SC04] se puede encontrar una revisión muy completa y detallada de las funciones de *kernel* que se han propuesto durante los últimos años en los distintos campos de aplicación en el ámbito del aprendizaje máquina. De todas ellas, las funciones de *kernel* polinómico, sigmoidal, Gaussiano, ANOVA, etc., son algunos de los ejemplos más representativos.

Llegados a este punto, se dispone ya de la base teórica necesaria para introducir la formulación de la máquina de vectores soporte en su modalidad para clasificación.

3.2. Máquina de vectores soporte para clasificación: SVC

3.2.1. Formulación

La SVC es un clasificador binario que asigna una etiqueta $y \in \{+1, -1\}$ a cada una de las muestras de entrada $\mathbf{x}$ conforme al signo de la siguiente expresión:

$$f(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}) + b \tag{3.8}$$

donde $\phi : R^d \rightarrow H$ es una transformación del espacio de entrada a un espacio de características de igual o mayor dimensión (incluso infinita), en el que se supone una mayor separación entre las clases. El vector $\mathbf{w}$ define el hiperplano de decisión en dicho espacio y b representa el sesgo respecto al origen de coordenadas.

La máquina de vectores soporte es una generalización no lineal del hiperplano óptimo de separación para problemas no separables, por lo que la formulación de la SVC parte del funcional (3.3). La SVC aborda el problema de clasificación no separable, relajando el concepto de margen y permitiendo errores en la clasificación, para lo que se introducen unas variables $\xi_i \geq 0 \ (i = 1, ..., n)$ en la restricción (3.4) que representan el error que se comete en cada muestra de entrenamiento. Por lo tanto, la SVC queda formulada como el siguiente problema de minimización cuadrática:

$$\min_{\mathbf{w}, b, \xi_i} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{n} \xi_i \tag{3.9}$$

$$\text{sujeto a } y_i \left(\mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_i) + b \right) \geq 1 - \xi_i; \ \forall i = 1, ..., n \tag{3.10}$$

$$\xi_i \geq 0;\ (i = 1,...,n) \tag{3.11}$$

donde $\mathbf{x}_i \in R^d\ (i = 1,...,n)$ son las muestras de entrenamiento con etiquetas $y_i \in \{+1,-1\}$, y $C > 0$ es el parámetro que determina el balance entre maximizar el margen y minimizar el error de entrenamiento; a mayor valor de C, se centra en minimizar el error de entrenamiento; cuanto más pequeño, el objetivo será maximizar el margen.

El problema de optimización planteado en (3.9) se transforma, introduciendo los multiplicadores de Lagrange, en el Langrangiano:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\alpha_i\left[y_i\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i) + b\right) - 1 + \xi_i\right] - \sum_{i=1}^{n}\mu_i\xi_i \tag{3.12}$$

donde $\alpha_i, \mu_i \geq 0$ son los multiplicadores de Lagrange. A partir de este Langrangiano, mediante el planteamiento de las condiciones KKT (Karush-Kuhn-Tucker) (véase [KT51][Bur98], pág. 131), se llega al dual de Wolfe [NW99], que debe ser maximizado respecto a los multiplicadores de Lagrange α_i. Así, la SVC queda formulada como el siguiente problema de maximización:

$$\max_{\alpha_i} L = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{\phi}^T(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_j) \tag{3.13}$$

$$\textit{sujeto a}\ \sum_{i=1}^{n}\alpha_i y_i = 0; \tag{3.14}$$

$$0 \leq \alpha_i \leq C;\ \forall i = 1,...,n \tag{3.15}$$

Este problema de optimización es un problema cuadrático convexo, por lo que converge a una solución óptima global, donde la solución está dada por: el vector director del hiperplano de separación $\mathbf{w}$, que admite una expansión en términos de los vectores de entrenamiento en el espacio transformado de la forma:

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{\phi}(\mathbf{x}_i) \tag{3.16}$$

donde solo aquellas muestras cuyo multiplicador asociado α_i es distinto de 0 contribuyen a la definición de la frontera de decisión, razón por la que reciben el nombre de vectores soporte, y el parámetro b, que se calcula mediante el uso de las condiciones KKT $y_i\left(\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)+b\right)-1=0;\ \forall_i=1,...,n$.

Normalmente, la función $\boldsymbol{\phi}(\cdot)$ no se conoce de forma explícita o es imposible de evaluar. No obstante, el problema de optimización (3.13) únicamente precisa calcular los productos escalares $\phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$, los cuales se pueden evaluar mediante la función de *kernel* $K(\mathbf{x}_i,\mathbf{x}_j)$. Aunque, en general, el vector de pesos $\mathbf{w}$ no podrá calcularse, sustituyendo su expresión en (3.8) se llega a la salida blanda de la SVC:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i y_i K(\mathbf{x}_i,\mathbf{x}) + b \tag{3.17}$$

La complejidad computacional de la SVC aumenta con el número de muestras de entrenamiento, lo que dificulta su aplicación a grandes conjuntos de datos.

3.2.2. Least Squares SVC

El método de mínimos cuadrados (*Least Squares*, LS) SVC, propuesto por Suykens y Vandewalle [SV99], es una reformulación de las SVMs de Vapnik en el que la optimización consiste en resolver un sistema de ecuaciones lineales más sencillo de utilizar que los métodos tradicionales de entrenamiento basados en técnicas de programación cuadrática (*Quadratic Programming*, QP), pudiendo tratar con una cantidad más considerables de datos, que converge a la solución de la SVM [PNR+99].

A continuación se presenta una breve revisión del algoritmo LS-SVC. El procedimiento LS-SVC se obtiene mediante la reformulación de la SVC planteando el problema de minimización siguiente:

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 \tag{3.18}$$

$$\text{sujeto a } y_i - \left(\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right) + b\right) = \xi_i;\ \forall i = 1,...,n \tag{3.19}$$

Introduciendo los multiplicadores de Lagrange, se obtiene el siguiente Langrangiano:

$$L = \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{2}\sum_{i=1}^{n}\xi_i^2 - \sum_{i=1}^{n}\beta_i\left[\left(\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right) + b\right) + \xi_i - y_i\right] \tag{3.20}$$

donde $\beta_i \geq 0$ son los multiplicadores de Lagrange. Aplicando las condiciones de optimalidad (condiciones KKT) y eliminando $\mathbf{w}$ y ξ_i, se obtiene un sistema lineal en vez de un problema de optimización de programación cuadrática:

$$\begin{bmatrix} \mathbf{K} + \frac{1}{C}\mathbf{I} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{3.21}$$

donde $\mathbf{y} = \left[y_1, y_2,..., y_n\right]^T$, $\mathbf{1} = \left[1_1, 1_2,..., 1_n\right]^T$ y $\boldsymbol{\beta} = \left[\beta_1, \beta_2,..., \beta_n\right]^T$. $\mathbf{I}$ es la matriz identidad de tamaño nxn y $\mathbf{K}$ es la matriz de *kernels*, definida como $\mathbf{K}_{ij} = K\left(\mathbf{x}_i, \mathbf{x}_j\right)\ \forall i, j = 1,..., n.$

A partir de la solución $\left[\boldsymbol{\beta}\ b\right]^T$ de este sistema de ecuaciones se obtiene la salida blanda de la LS-SVC:

$$f(\mathbf{x}) = \sum_{i=1}^{n} \beta_i K(\mathbf{x}_i, \mathbf{x}) + b \tag{3.22}$$

donde debe señalarse que β_i converge de manera asintótica al término $\alpha_i y_i$ en (3.17) y, por lo tanto, la LS-SVC converge a la SVC original.

La desventaja de este modelo, respecto a la SVC, es que no se obtiene una solución dispersa en términos del número de vectores soporte, que puede influir en su capacidad de generalización.

3.2.3. One-class SVC

La clasificación *one-class*, es un tipo especial de problema de clasificación de dos clases, donde cada una de las clases tiene un significado específico: la clase *target* y la clase *outlier* o clase anómala. Se asume que las muestras de la clase *target* son bien conocidas, mientras que las muestras de la clase *outlier*, pueden ser escasas o ninguna [TD04].

Schölkopf en [SWS+00] propone un enfoque que es llamado *v-Support Vector Classification* (*v*-SVC) que utiliza un hiperplano para separar las muestras objetivo desde el origen con un margen máximo. Tax y Duin en [TD04] propusieron otra alternativa que denominaron *Support Vector Data Description* (SVDD), y es dicha solución la que se implementa en este proyecto. Para describir el dominio del conjunto de datos, se puede considerar que todos los elementos del mismo, quedan encerrados en una hiperesfera con volumen mínimo. Minimizando el volumen del espacio de características capturado, se espera minimizar la posibilidad de aceptar objetos *outliers* dentro de la hiperesfera. Cuando todos los datos están normalizados a la unidad, el modelo SVDD es equivalente al *v*-SVC. El modelo SVDD es un caso especial de máquina de vectores soporte para clasificación [Vap95].

El clasificador SVDD define una hiperesfera caracterizada por su centro $\mathbf{a} \in R^d$ y radio $R \in R$ que contiene a todas las muestras de entrenamiento en el espacio transformado $\boldsymbol{\phi}(\mathbf{x}_i)$. La función error que hay que minimizar es:

$$\min_{R,\mathbf{a},\xi_i} F(R,\mathbf{a}) = R^2 + C\sum_{i=1}^{n}\xi_i \tag{3.23}$$

$$\textit{sujeto a } \left\|\boldsymbol{\phi}(\mathbf{x}_i) - \mathbf{a}\right\|^2 \le R^2 + \xi_i;\ \forall i = 1,...,n \tag{3.24}$$

$$\xi_i \ge 0\ \ (i = 1,...,n) \tag{3.25}$$

donde el parámetro C se introduce para controlar el balance entre el volumen de la hiperesfera y el número de errores en el entrenamiento. El functional (3.23) se transforma, introduciendo los multiplicadores de Lagrange, en el Langrangiano:

$$L = R^2 + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left[\left\|\boldsymbol{\phi}(\mathbf{x}_i) - \mathbf{a}\right\|^2 - R^2 - \xi_i\right] - \sum_{i=1}^{n}\mu_i\xi_i \tag{3.26}$$

donde $\alpha_i, \mu_i \ge 0$ son los multiplicadores de Lagrange. Se llega a una expresión dual, expresada según la función de *kernel*, que debe ser minimizada respecto a los multiplicadores de Lagrange α_i. Así, la SVDD queda formulada como el siguiente problema de minimización:

$$\min_{\alpha_i} L = \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j k(\mathbf{x}_i,\mathbf{x}_j) - \sum_{i=1}^{n}\alpha_i k(\mathbf{x}_i,\mathbf{x}_i) \tag{3.27}$$

$$\textit{sujeto a } \sum_{i=1}^{n}\alpha_i = 1; \tag{3.28}$$

$$0 \le \alpha_i \le C;\ \forall i = 1,...,n \tag{3.29}$$

Método de entrenamiento basado en técnicas de programación cuadrática, donde la solución está dada por: el centro de la hiperesfera $\mathbf{a}$, que admite una expansión en términos de los vectores de entrenamiento en el espacio transformado de la forma:

$$\mathbf{a} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\mathbf{x}_i) \tag{3.30}$$

donde solo aquellas muestras cuyo multiplicador asociado α_i es distinto de 0 contribuyen a la definición de la descripción del conjunto de datos, razón por la que reciben el nombre de vectores soporte, y el radio de la hiperespera R, que se obtiene usando cualquiera de los vectores soporte $\mathbf{x}'$ de la forma:

$$R^2 = k(\mathbf{x}', \mathbf{x}') - 2\sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}') + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{3.31}$$

Como resultado de salida del sistema, se obtiene la función de decisión siguiente:

$$f(\mathbf{x}) = R^2 - k(\mathbf{x}, \mathbf{x}) + 2\sum_{i=1}^{n} \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j) \tag{3.32}$$

donde las muestras con mayor valor son más parecidas a las muestras de la clase *target*.

Capítulo 4

4. Desarrollo del proyecto

A continuación se explican en detalle, descritas en el Capítulo 1, las diferentes fases en las que se ha desarrollado el proyecto. Además se muestran los diferentes problemas y las respectivas decisiones tomadas para la solución de este proyecto.

4.1. Estudio y tratamiento de la base de datos

Las bases de datos biomédicas contienen generalmente variables de características de alta dimensionalidad y múltiples tipos de clases. Los datos obtenidos de las bases de datos contienen errores sistemáticos y errores humanos. La presencia de ruido (identificación de datos incompletos, incorrectos, inexactos, no pertinentes, etc.) y la ausencia de valores, obstaculizan la exactitud de los sistemas de aprendizaje máquina. Estas inconsistencias hacen tratar el problema de clasificación de conjuntos de datos biomédicos como un caso especial [Lav99].

En este caso, la base de datos proporcionada, en Excel, se trata de un histórico de datos de 1100 pacientes de la UCI del Hospital Clínico Universitario de Valladolid, debidamente procesado para preservar el anonimato de dichos pacientes. Se tienen 33 variables, tanto binarias como cualitativas y cuantitativas, relacionadas con la evolución del paciente en el servicio de la UCI. En la *Tabla 2*, se muestra un resumen de las características presentes en la base de datos a estudio.

Instancias	**Variables**			**Datos Ausentes (%)**	**Ruido (%)**
	Cuantitativas discretas	**Binarias**	**Nominales**		
1100	9	4	20	0.64	17.64

Tabla 2. Resumen de características la base de datos

Existen diferentes técnicas para tratar de resolver el problema de los datos ausentes y el ruido. Una de ellas, que es la que se ha usado en este proyecto, es descartar las muestras que tengan falta o error en alguna de las variables [TK03]. Se ha descartado cualquier otra técnica ya que no es objeto de estudio de este proyecto, considerando que la pérdida de información no era demasiado significativa. Para la limpieza de datos, mediante el estudio de los registros de la base de datos proporcionada, se han eliminado 201 entradas, 7 por falta de datos y 194 debido a algún error en los datos, quedando así 899 entradas de pacientes para el estudio, y cada entrada en la base de datos corresponde con una muestra, ya que no se tienen en cuenta los reingresos de los pacientes en la UCI.

Una vez eliminados los datos, se importa la base de datos a Matlab para posteriormente trabajar mejor la parte de la algoritmia.

4.2. Elección, codificación y estudio de las variables de interés

Se han seleccionado 14 de las 33 variables existentes, debido a su significado para la evolución de los pacientes en el servicio de la UCI. En la *Tabla 3* se muestran las variables seleccionadas, y su significado.

Variable	Significado
FechaNacimiento	Fecha de nacimiento.
Sexo	Sexo del paciente, hombre o mujer.
FechaIngresoHospital	Día de ingreso en el hospital.
FechaIngresoUCI	Día de ingreso en la UCI.
Procedencia	Área de origen de llegada al hospital. Pacientes procedentes de otro servicio del mismo hospital, pacientes de otros hospitales, o urgencias.
CausaIngresoHospital	Motivo por el que el paciente ingresa en el hospital.
Enfermedad	Enfermedad que presenta el paciente.
APACHEII	Índice que predice la evolución de los pacientes en el servicio de la UCI.
Antecedentes	Antecedentes que presenta el paciente.
Diagnósticos	Diagnósticos realizados por el equipo médico.
Técnicas	Procedimientos realizados por el equipo médico al paciente en el servicio de la UCI.
Complicaciones	Complicaciones que le han surgido al paciente en el servicio de la UCI.
FechaAltaUCI	Día que se le da el alta al paciente.
MotivoAlta	Indica si la evolución del paciente ha sido favorable o no.

Tabla 3. Variables seleccionadas para el estudio

A partir de las variables anteriormente descritas se crea el conjunto de datos con el que se va a trabajar posteriormente, como se describe a continuación:

- ***Edad***. Variable cuantitativa discreta. Conocida la fecha de nacimiento, se ha calculado la edad (en años) del paciente como un número natural. Se tienen pacientes con una edad comprendida entre los 8 y 82 años, con una media aritmética de 60.85 años, y una desviación estándar de 16.52 años. En la *Figura 3* se muestra la distribución de la edad de los pacientes de la UCI, donde se observa que existe un segmento significativo de pacientes de avanzada edad.

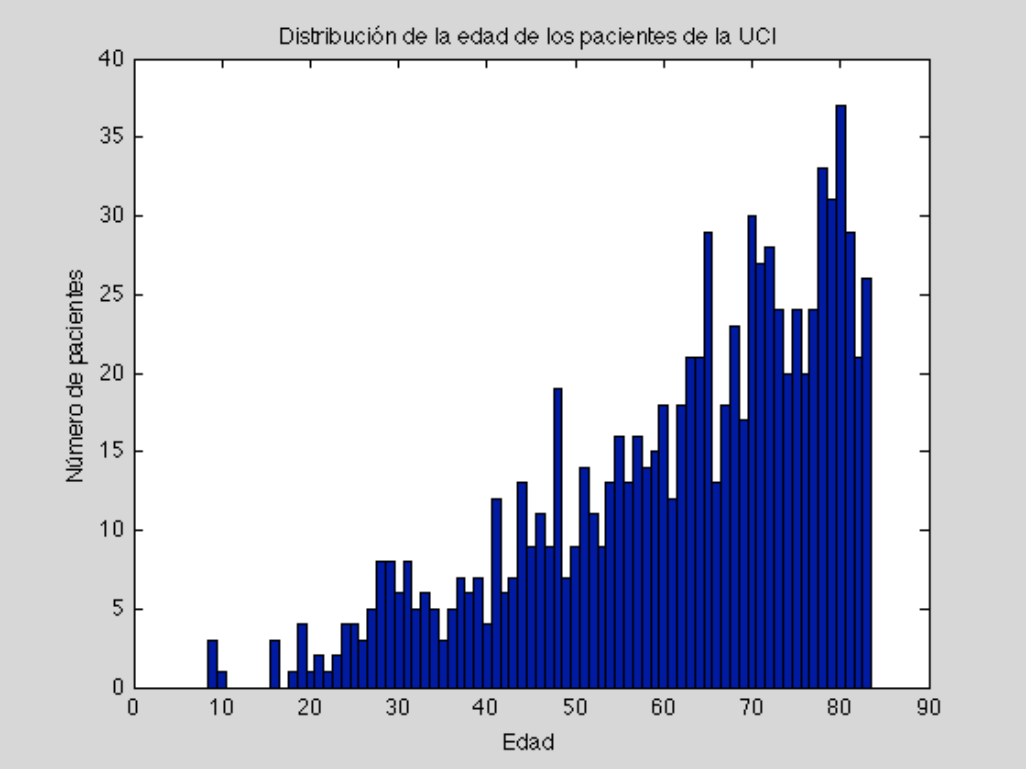

Figura 3. Distribución de la edad de los pacientes de la UCI

- ***Sexo***. Variable cualitativa nominal, donde se distingue si el paciente es hombre o mujer. En caso de ser hombre se codifica con un "1", y si es mujer con un "0". En la *Tabla 4* se muestra la transformación utilizada para la variable *Sexo*.

Sexo	
Categoría	**Clase**
Hombre	1
Mujer	0

Tabla 4. Transformación de la variable Sexo

Los porcentajes de hombres y mujeres, respecto al número total de pacientes, están repartidos de la forma siguiente: el 68.41% de pacientes son hombres y el 31.59% son mujeres, como se muestra en la *Figura 4*.

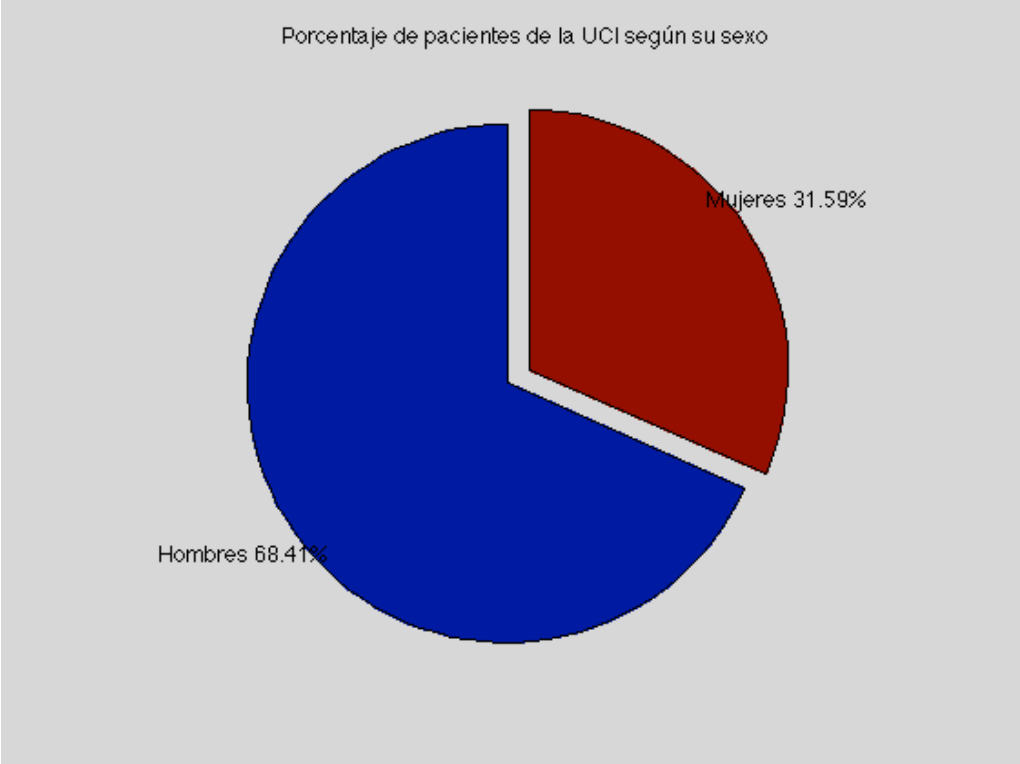

Figura 4. Porcentaje de pacientes de la UCI según su sexo

- ***Días transcurridos hasta el ingreso en UCI***. Variable cuantitativa discreta. Son los días que pasan desde que el paciente ingresa en el hospital hasta que es ingresado en el servicio de la UCI. Días codificados como un número natural. Los pacientes pasan entre 0 y 94 días ingresados en el hospital hasta que son

ingresados en el servicio de la UCI, con una media aritmética de 6.3 días, y una desviación estándar de 11.42 días. En la *Figura 5* se muestra la distribución de los días transcurridos hasta el ingreso en UCI, donde se observa que aproximadamente el 40% de los pacientes se ingresan en la UCI el mismo día que son hospitalizados.

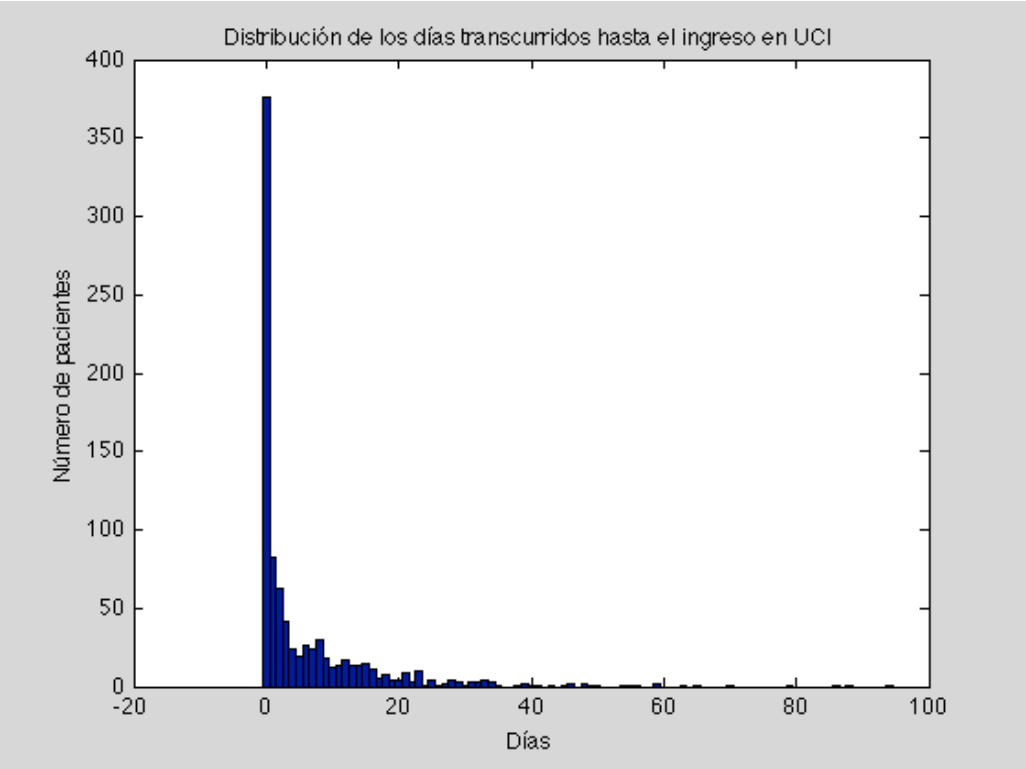

Figura 5. Distribución de los días transcurridos hasta el ingreso en UCI

- ***Procedencia***. Variable cualitativa nominal. Indica el área de origen de llegada del paciente al hospital. Para pacientes del mismo hospital, procedentes de otros servicios, se codifica con un "1", en pacientes procedentes de otros hospitales, se codifica con un "2", y para aquellos que llegan de servicios de urgencias, se codifica con un "3". En la *Tabla 5* se muestra la transformación utilizada para la variable *Procedencia*.

Procedencia	
Categoría	**Clase**
Mismo hospital	1
Otros hospitales	2
Urgencias	3

Tabla 5. Transformación de la variable Procedencia

Se ha observado que el 82.09% de pacientes son del mismo hospital, el 13.79% proceden de otros hospitales, y el 4.12% de servicios de urgencias, como se muestra en la *Figura 6*.

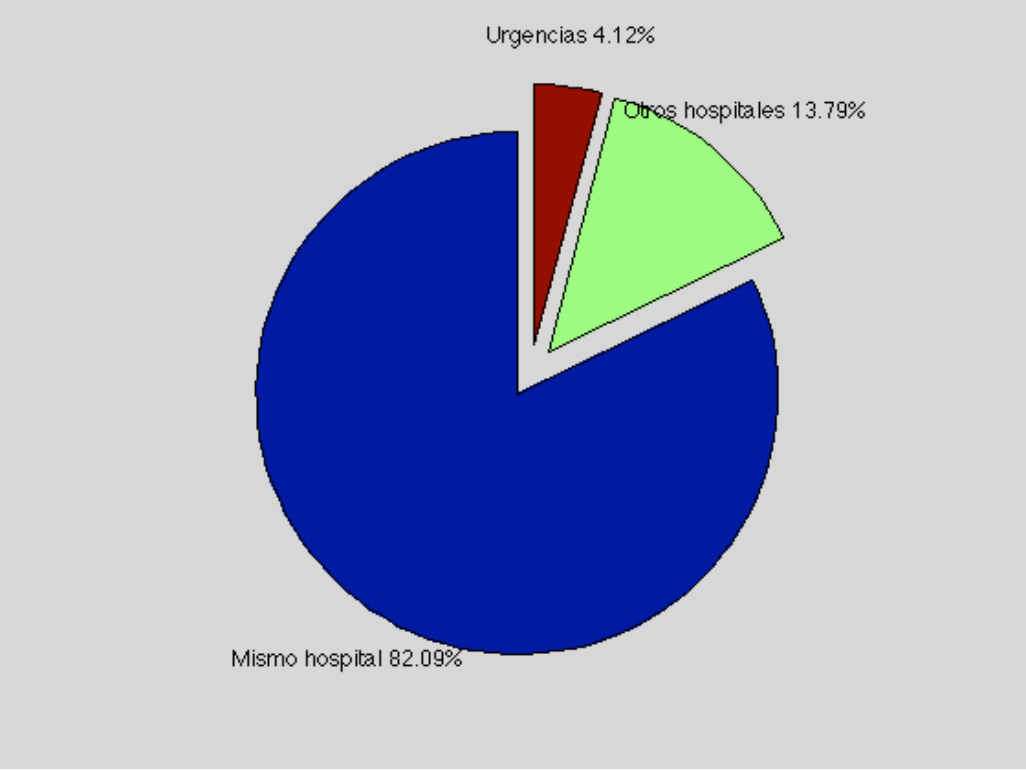

Figura 6. Porcentaje de pacientes de la UCI según su procedencia

- ***Causa de ingreso al hospital***. Variable cualitativa nominal. Indica el motivo por el que el paciente ingresa en el hospital. Se trata de un código de 4 cifras, de valor comprendido entre el número 1000 y el número 9999, que definimos como

código enfermedad. El primer dígito indica el aparato (circulatorio, respiratorio, etc.), el segundo dígito denota el grupo perteneciente, y los dos últimos dígitos es la especificación correspondiente. Para codificar esta variable se han utilizado tres campos, donde el primer campo, es el valor del primer dígito del código enfermedad, el segundo campo, es el valor formado por el primer y segundo dígito del código enfermedad, y el tercer campo, coincide con el código enfermedad. La codificación de esta variable forma un vector de dimensión 3.

A continuación se muestra un ejemplo de esta representación para el código enfermedad 7542, código de 4 cifras donde el primer dígito, de valor 7, indica el aparato, el segundo dígito, de valor 5, indica el grupo perteneciente, y los dos últimos dígitos, de valor 42, indican la especificación correspondiente:

7	75	7542
Primer campo	Segundo campo	Tercer campo

En la *Tabla 6* se muestran las diez causas de ingreso al hospital que más se repiten entre los pacientes de la UCI.

Código	Causa de ingreso al hospital	Número de pacientes	Porcentaje
3820	CIRUGÍA ENFERMEDAD MALIGNA DE TRÁQUEA, BRONQUIOS Y PULMON	112	12.46%
4950	POLITRAUMATIZADO	61	6.79%
3800	CIRUGÍA DE PARED TORÁCICA	53	5.9%
6230	ACCIDENTE CEREBROVASCULAR HEMORRÁGICO	44	4.89%
3810	CIRUGÍA ENFERMEDAD BENIGNA DE TRÁQUEA, BRONQUIOS Y PULMÓN	31	3.45%
4113	TRAUMATISMO CRÁNEOENCEFÁLICO GRAVE GCS 3-8	24	2.67%

2300	INSUFICIENCIA RESPIRATORIA CRÓNICA AGUDIZADA	23	2.56%
6260	HEMORRAGIA SUBARACNOIDEA NO TRAUMÁTICA	18	2%
2132	INSUFICIENCIA RENAL AGUDA SECUNDARÍA A INFECCIÓN DE VÍAS BAJAS, EXTRAHOSPITALARIA	13	1.45%
4961		13	1.45%

Tabla 6. Causas de ingreso al hospital más frecuentes entre los pacientes de la UCI

- ***Enfermedad***. Variable cualitativa nominal. Enfermedad que tiene el paciente, representada con un código enfermedad, por lo tanto, su codificación es la misma que para la variable *Causa de ingreso al hospital*. En la *Tabla 7* se muestran las diez enfermedades que más se repiten entre los pacientes de la UCI.

Código	Enfermedad	Número de pacientes	Porcentaje
3820	CIRUGÍA ENFERMEDAD MALIGNA DE TRÁQUEA, BRONQUIOS Y PULMON	137	15.24%
4950	POLITRAUMATIZADO	57	6.34%
3810	CIRUGÍA ENFERMEDAD BENIGNA DE TRÁQUEA, BRONQUIOS Y PULMÓN	51	5.67%
6230	ACCIDENTE CEREBROVASCULAR HEMORRÁGICO	44	4.89%
3800	CIRUGÍA DE PARED TORÁCICA	36	4%
1830	SHOCK SÉPTICO	24	2.67%
4113	TRAUMATISMO CRÁNEOENCEFÁLICO GRAVE GCS 3-8	23	2.56%
2300	INSUFICIENCIA RESPIRATORIA CRÓNICA AGUDIZADA	21	2.34%

6260	HEMORRAGIA SUBARACNOIDEA NO TRAUMÁTICA	17	1.89%
6871	COMA DE DIVERSO ORIGEN	17	1.89%

Tabla 7. Enfermedades más frecuentes entre los pacientes de la UCI

- ***APACHE II***. Variable cuantitativa discreta. Número natural que predice la evolución de los pacientes en el servicio de la UCI. Dicha cuantificación está comprendida entre los valores 0 y 49, con una media aritmética de 13.62, y una desviación estándar de 8.96. En la *Figura 7* se muestra la distribución del índice APACHE II de los pacientes de la UCI.

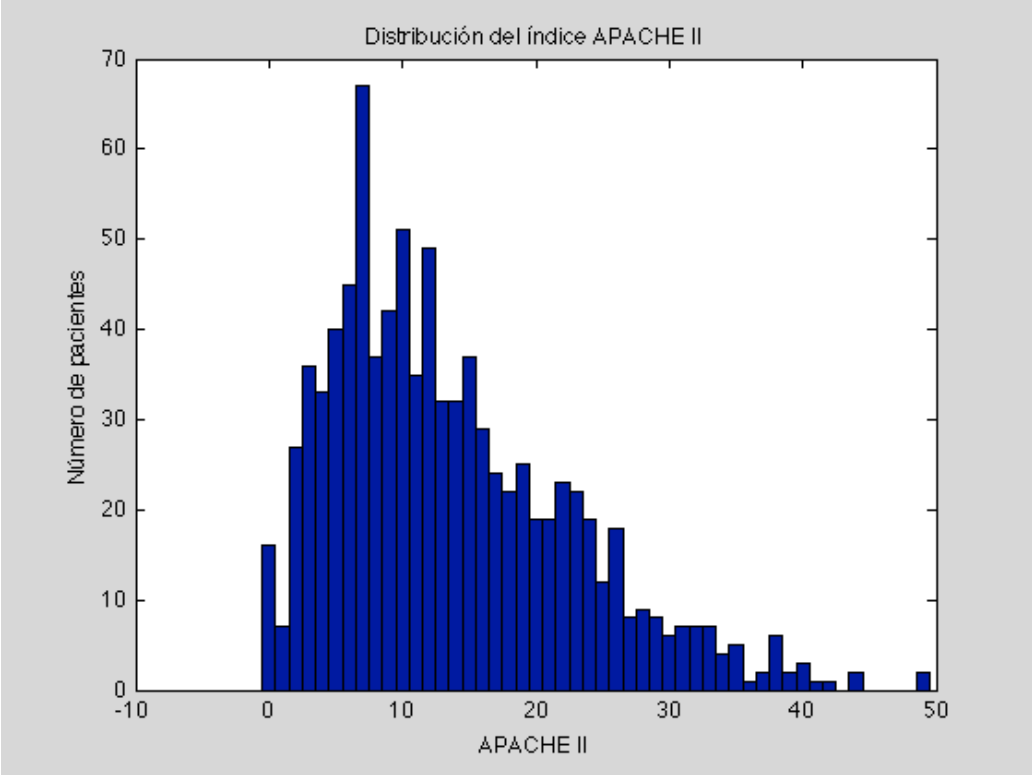

Figura 7. Distribución del índice APACHE II de los pacientes de la UCI

- ***Antecedentes***. Variable cualitativa nominal no mutuamente exclusiva, es decir, cada paciente puede tener diferentes antecedentes. Antecedentes que presenta el paciente. Se tratan de códigos enfermedad, pero en este caso como cada paciente

puede tener diferentes antecedentes, para la codificación de esta variable se ha utilizado una codificación "1-de-C" con tres campos. En el primer campo, de dimensión igual a 9, se acumulan las unidades que coinciden con el valor formado por el primer dígito de los códigos enfermedad, en el segundo campo, de dimensión igual a 99, se acumulan las unidades que coinciden con el valor formado por el primer y segundo dígito de los códigos enfermedad, y en el tercer campo, de dimensión igual a 9999, se acumulan las unidades que coinciden con el valor formado por los códigos enfermedad. La codificación de esta variable forma un vector de dimensión 9+99+9999.

A continuación se muestra un ejemplo de esta representación para un paciente que tiene como antecedentes los códigos enfermedad 7520 y 7400.

	0	0	0	0	0	0	2	0	0	0	··	1	1	··	0	0	··	1	··	1	··	0
Posición	1	2	3	4	5	6	7	8	9	1	···	74	75	···	99	1	···	7400	···	7520	···	9999
	Primer campo											Segundo campo						Tercer campo				

Para el primer código enfermedad, el primer dígito se codifica en el primer campo activando la unidad de la posición 7, el valor formado por el primer y segundo dígito se codifica en el segundo campo activando la unidad de la posición 75, y el código enfermedad se codifica en el tercer campo activando la unidad de la posición 7520. Con el segundo código enfermedad, el primer dígito coincide con el primer dígito del primer código enfermedad y la unidad de la posición 7 se actualiza al valor 2, el valor formado por el primer y segundo dígito se codifica en el segundo campo activando la unidad de la posición 74, y el código enfermedad se codifica en el tercer campo activando la unidad de la posición 7400. En el caso de que el paciente tuviera más códigos de enfermedad como antecedentes, se procedería de igual forma a la descrita anteriormente. En la *Tabla 8* se muestran los diez antecedentes que más se repiten entre los pacientes de la UCI.

Código	Antecedentes	Número de pacientes	Porcentaje
7991	TABAQUISMO	189	21.02%
1710	HIPERTENSIÓN ARTERIAL NO DESCOMPENSADA	178	19.8%
7610	DIABETES MELLITUS TIPO 1	63	7.01%
7910	ALCOHOLISMO CRÓNICO	53	5.9%
7940	HIPERLIPIDEMIAS	47	5.23%
1364	FIBRILACIÓN AURICULAR	42	4.67%
7992	ALERGIAS	41	4.56%
3320	CIRUGÍA ENFERMEDAD BENIGNA GASTRODUODENAL. HEMORRAGÍA DIGESTIVA ALTA	40	4.45%
3500	UROLOGIA. CIRUGÍA COMPLEMENTARIA Y ENFERMEDADES UROLÓGICAS MÉDICAS	38	4.23%
3300	CIRUGIA DIGESTIVA/ABDOMINAL	30	3.34%

Tabla 8. Antecedentes más frecuentes entre los pacientes de la UCI

- ***Diagnósticos***. Variable cualitativa nominal no mutuamente exclusiva, es decir, a cada paciente se le han podido realizar diferentes diagnósticos. Diagnóstico realizado por el equipo médico. Se tratan de códigos enfermedad, y por tanto, la codificación de esta variable para cada paciente es la misma que para la variable *Antecedentes*. En la *Tabla 9* se muestran los diez diagnósticos que más se repiten entre los pacientes de la UCI.

Código	Diagnóstico	Número de pacientes	Porcentaje
3820	CIRUGÍA ENFERMEDAD MALIGNA DE TRÁQUEA, BRONQUIOS Y PULMON	140	15.57%
3810	CIRUGÍA ENFERMEDAD BENIGNA DE TRÁQUEA, BRONQUIOS Y PULMÓN	51	5.67%
4300	TRAUMATISMO TORÁCICO	46	5.12%
1830	SHOCK SÉPTICO	44	4.89%
6230	ACCIDENTE CEREBROVASCULAR HEMORRÁGICO	41	4.56%
3800	CIRUGÍA DE PARED TORÁCICA	36	4%
4113	TCE GRAVE GCS 3-8	36	4%
4950	POLITRAUMATIZADO	31	3.45%
4720	TRAUMATISMO DE HOMBRO, ESCÁPULA, CODO, HÚMERO Y ANTEBRAZO	29	3.23%
4361	CONTUSIÓN Y HEMORRAGIA PULMONAR	27	3%

Tabla 9. Diagnósticos más frecuentes entre los pacientes de la UCI

- ***Técnicas***. Variable cualitativa nominal no mutuamente exclusiva, es decir, a cada paciente se le han podido realizar diferentes técnicas. Procedimientos realizados por el equipo médico al paciente en el servicio de la UCI. Se trata de un código de 4 cifras, de valor comprendido entre el número 9100 y el número 9999, que definimos como **código técnica**. Comienza por 9, y luego el segundo dígito indica el aparato sobre el que se actúa, y los dos últimos dígitos es la técnica en general. Se ha utilizado una codificación "1-de-C" con dos campos, para codificar esta variable de manera similar a las variables *Antecedentes* y *Diagnósticos*, como se indica a continuación. En el primer campo, de dimensión igual a 9, se acumula la unidad que coincide con el valor del segundo dígito de los códigos técnica, y en el

segundo campo, de dimensión igual a 999, se acumula la unidad que coincide con el valor formado del segundo, tercer, y cuarto dígito de los códigos técnica. La codificación de esta variable forma un vector de dimensión 9+999.

A continuación se muestra un ejemplo de esta representación para un paciente que tiene como técnicas los códigos técnica 9618 y 9640.

											··		··		··	
	0	0	0	0	0	2	0	0	0	0	··	1	··	1	··	0
Posición	1	2	3	4	5	6	7	8	9	1	···	618	···	640	···	999
	Primer campo											Segundo campo				

Para el primer código técnica, el segundo dígito se codifica en el primer campo activando la unidad de la posición 6, el valor formado por el segundo, tercer, y cuarto dígito se codifica en el segundo campo activando la unidad de la posición 618. Con el segundo código técnica, el segundo dígito coincide con el segundo dígito del primer código técnica y la unidad de la posición 6 se actualiza al valor 2, el valor formado por el segundo, tercer, y cuarto dígito se codifica en el segundo campo activando la unidad de la posición 640. En el caso de que el paciente tuviera más códigos técnica, se procedería de igual forma a la descrita anteriormente. En la *Tabla 10* se muestran las diez técnicas más utilizadas por los especialistas entre los pacientes.

Código	**Técnica**	**Número de pacientes**	**Porcentaje**
9609	YUGULAR INTERNA	323	35.93%
9211	INTUBACIÓN OROTRAQUEAL	302	33.59%
9222	CMV+PEEP	237	26.36%
9221	CMV	225	25.03%
9641	SEDACIÓN EN PERFUSIÓN CONTÍNUA	200	22.25%
9637	ANALGESIA INTRAVENOSA	161	17.91%
9634	ANTICOAGULACIÓN CRÓNICA. PROFILAXIS ANTITROMBÓTICA	156	17.35%

9517	TAC CEREBRAL	144	16.02%
9408	BRONCOASPIRADOS	125	13.9%
9407	HEMOCULTIVOS	117	13.01%

Tabla 10. Técnicas más frecuentes entre los pacientes de la UCI

- ***Complicaciones***. Variable cualitativa nominal no mutuamente exclusiva, es decir, a cada paciente se le han podido detectar diferentes complicaciones. Complicaciones que le han surgido al paciente en el servicio de la UCI. Se tratan de códigos enfermedad, y por tanto, la codificación de esta variable para cada paciente es la misma que para las variables *Antecedentes* y *Diagnósticos*. En la *Tabla 11* se muestran las diez complicaciones que más se repiten entre los pacientes de la UCI.

Código	Complicación	Número de pacientes	Porcentaje
6880	EVOLUCION ROSTROCAUDAL Y MUERTE CEREBRAL	41	4.56%
5940	FRACASO MULTIORGÁNICO	36	4%
1392	ASISTOLIA (PARO CARDIO RESPIRATORIO) SECUNDARIA	29	3.23%
2560	ATELECTASIA	24	2.67%
1364	FIBRILACIÓN AURICULAR	16	1.78%
1830	SHOCK SÉPTICO	15	1.67%
5411	NEUMONÍA COMUNITARIA	15	1.67%
5930	SEPTICEMIA/BACTERIEMIA PRIMARIA	14	1.56%
6810	DESORIENTACION/DELIRIO AGUDO	14	1.56%
1720	HIPERTENSION ARTERIAL	12	1.33%

Tabla 11. Complicaciones más frecuentes entre los pacientes de la UCI

- ***Días en el servicio de la UCI***. Variable cuantitativa discreta. Son los días que pasa el paciente ingresado en el servicio de la UCI, desde que ingresa en el servicio la UCI, hasta que se le concede el alta. Días codificados como un número natural. Los pacientes pasan entre 0 y 345 días ingresados en el servicio de la UCI, con una media aritmética de 6.72 días, y una desviación estándar de 16.18 días. En la *Figura 8* se muestra la distribución de los días en el servicio de la UCI, donde se observa que aproximadamente el 75% de los pacientes permanecen ingresados en la UCI menos de una semana.

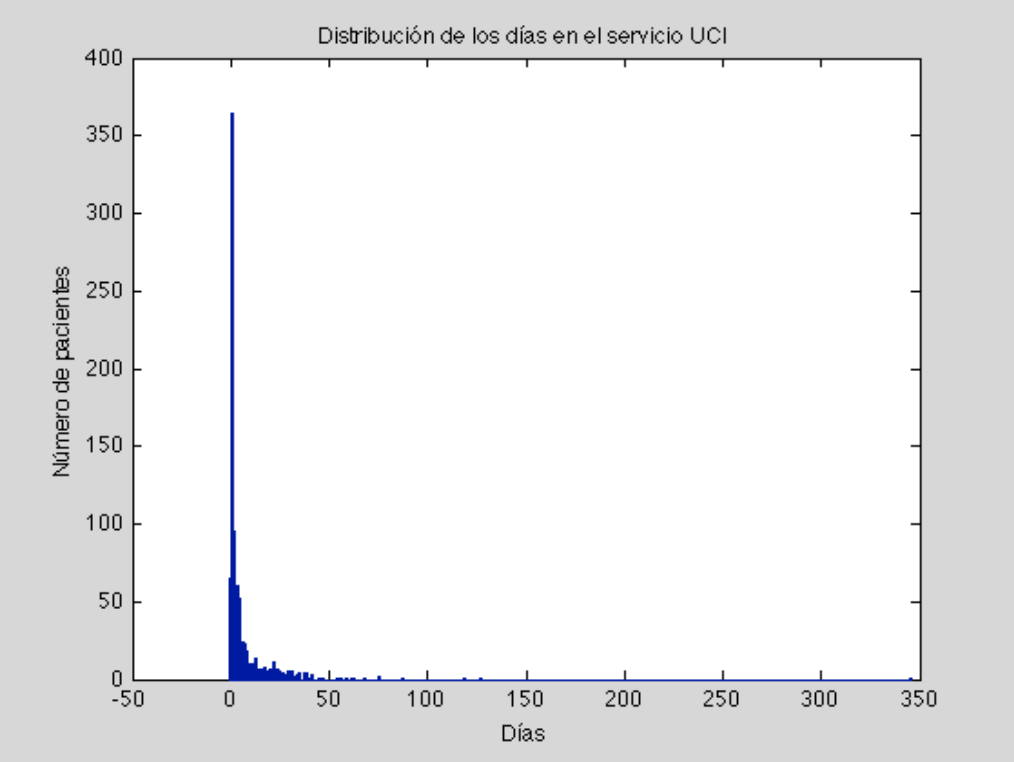

Figura 8. Distribución de los días en el servicio de la UCI

Motivo del alta. Variable cualitativa ordinal dicotómica. Se distingue si el paciente tiene una evolución favorable o muere en el servicio de la UCI. En caso de tener una evolución favorable se codifica con un "1", y en el caso contrario con un "0". %. En la *Tabla 12* se muestra la transformación utilizada para la variable *Motivo del alta*.

Motivo del alta	
Categoría	**Clase**
Vive	1
Muere	0

Tabla 12. Transformación de la variable Motivo del alta

El porcentaje de pacientes con evolución favorable es del 79.2% y el de pacientes que muere del 20.8, como se muestra en la *Figura 9*.

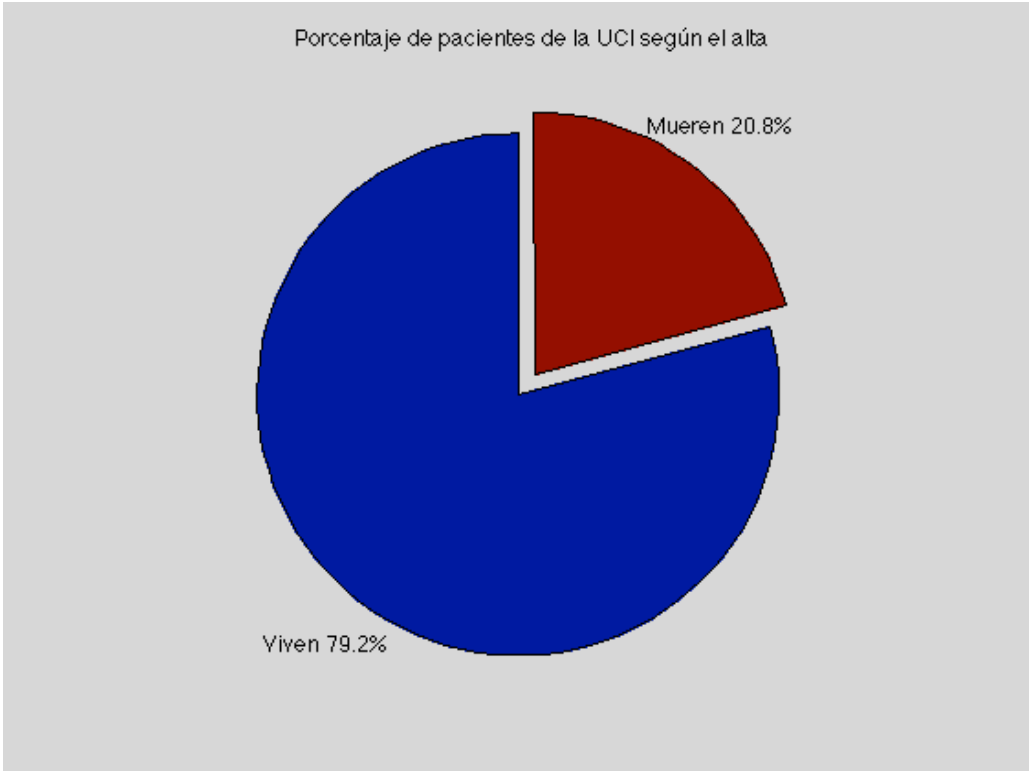

Figura 9. Porcentaje de pacientes de la UCI según el alta

Capítulo 4. Desarrollo del proyecto

Según el estudio realizado de las variables que forman el conjunto de datos, se pueden hacer las siguientes afirmaciones sobre los pacientes que ingresan en el servicio de la UCI:

- **La edad media es de 60.85 años**, donde se observa que existe un segmento significativo de pacientes de avanzada edad.
- **El doble de pacientes que ingresan en la UCI son hombres**. Concretamente el 68.41% corresponde a hombres y un 31.59% a mujeres.
- Aproximadamente el 40% de los pacientes se ingresan en la UCI el mismo día que son hospitalizados.
- La mayoría, el 82.09%, son pacientes procedentes de otros servicios del mismo hospital, el 13.79% son pacientes que proceden de otros hospitales y el 4.12% proceden de los servicios de urgencias.
- **La cirugía de enfermedad de TRÁQUEA, BRONQUIOS Y PULMÓN,** y el traumatismo**,** en particular**,** el **TRAUMATISMO TORÁCICO,** son las enfermedades diagnosticadas más frecuentes entre los pacientes que ingresan en la UCI.
- En media se tiene un **índice APACHE II de 13.62**.
- **Siendo el TABAQUISMO el antecedente común a aproximadamente el 20% de los pacientes**. Antecedente directamente relacionado con la cirugía de enfermedad de tráquea, bronquios y pulmón, una de las enfermedades diagnosticada más frecuente entre los pacientes, como se ha indicado anteriormente.
- **El 75% de los pacientes permanecen ingresados en la UCI menos de una semana**.
- Un 79.2% de los pacientes tienen una evolución favorable en el servicio de la UCI, mientras que **el 20.8% de pacientes muere**.

4.3. Aplicación de la SVM para la predicción de la mortalidad

Codificadas las variables de interés para el estudio, cada paciente estará caracterizado por un vector de dimensión $d = 31341$, definiendo así a X como el **conjunto de datos de entrada** o el conjunto de pacientes a estudiar**,** con $N = 830$ el número de pacientes considerados.

La máquina de vectores soporte, en su modalidad para clasificación (SVC), se utiliza para predecir a posteriori la mortalidad de cada uno de los pacientes de la UCI, es decir, saber si un paciente tiene una evolución favorable en el servicio de la UCI o, por el contrario, muere, comparando los resultados obtenidos mediante LS-SVC y SVDD.

En este proyecto, las clases consideradas por las SVMs se corresponden con la clase positiva, que son los pacientes que mueren en el servicio de la UCI con un determinado índice APACHE II bajo, cuya probabilidad de muerte es pequeña (pacientes etiquetados con +1), y la clase negativa, a la que pertenecen el resto de pacientes que no son de la clase positiva (pacientes etiquetados con -1). En términos de la *one-class* SVC, la clase positiva corresponde con la clase *outlier* y la clase negativa con la clase *target*.

El uso de la SVM mediante métodos *kernel*, *se* basa en la selección de una función de *kernel* adecuada, que depende en gran medida de las características del problema de aprendizaje que se aborda. El estudio de la influencia de la función de *kernel* empleada, sobre las prestaciones de las SVMs, queda fuera de los objetivos concretos de este proyecto, y se ha optado por emplear una función de *kernel* de tipo Gaussiano, que en general resulta suficientemente versátil:

$$K(\mathbf{x},\mathbf{x}') = e^{\frac{(D(\mathbf{x},\mathbf{x}'))^2}{2\sigma^2}} \tag{4.1}$$

donde $D(\mathbf{x},\mathbf{x}')$ es la distancia entre pacientes, y $\sigma > 0$ es la anchura de la función de base radial (4.1), siendo dicha anchura un parámetro a optimizar.

La adopción de una distancia entre pacientes, caracterizados por variables tanto binarias como cualitativas y cuantitativas, hace necesario el estudio de índices de similitud para un conjunto de variables mixtas. En este caso, se define el índice de similitud de Gower [Gow71] como:

$$S_G(\mathbf{x},\mathbf{x}') = \frac{\sum_{k=1}^{M} s_k(\mathbf{x},\mathbf{x}')}{\sum_{k=1}^{M} w_k} \tag{4.2}$$

donde $s_k(\mathbf{x},\mathbf{x}') = 1 - \frac{|x_k - x'_k|}{R_k}$, siendo R_k el rango de la k-ésima variable continua; $s_k(\mathbf{x},\mathbf{x}') = 1$ en el caso de coincidencia del tipo presencia-presencia para la k-ésima variable binaria; $s_k(\mathbf{x},\mathbf{x}') = 1$ en el caso de coincidencia para la k-ésima variable cualitativa, y finalmente w_k toma el valor 1 ó 0 dependiendo de si la comparación considerada es válida para la k-ésima variable. Para las variables binarias w_k toma el valor 0 ante una coincidencia del tipo ausencia-ausencia.

A partir del índice de similitud (4.2) se define la distancia de Gower como: $(D(\mathbf{x},\mathbf{x}'))^2 = 1 - S_G(\mathbf{x},\mathbf{x}')$. Dicha transformación no es la única posible del índice de similitud que ofrece una distancia, sin embargo, en este proyecto se ha utilizado precisamente ésta porque garantiza que la distancia obtenida es euclídea [Gow66].

Para construir las SVMs se define el conjunto de entrenamiento $T \subset X$, como el conjunto de pacientes prototipo. En el proceso de entrenamiento, la SVM se basa en la matriz de similitud $K(T,T)$, en relación a todos los pacientes. La información sobre un conjunto de *test* $S \subset X$, se proporciona en términos de sus similitudes con el conjunto T. En particular, $K(S,T)$ se trata de una descripción de un espacio donde cada dimensión se corresponde con la similitud a un paciente prototipo. En general, $K(\mathbf{x},T)$ define un vector que consiste en las similitudes existentes entre el paciente $\mathbf{x}$ y todos los pacientes del conjunto de entrenamiento T, es decir, si $T=\{\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_n\}$, entonces $K(\mathbf{x},T)=[K(\mathbf{x},\mathbf{x}_1),K(\mathbf{x},\mathbf{x}_2),...,K(\mathbf{x},\mathbf{x}_n)]^T$. Por la tanto, $K(\cdot,T)$ se puede entender como una aproximación de la función de transformación $\phi(\cdot)$, conocida como *Empirical Kernel Map* [SS02].

Para la LS-SVC, la clase positiva está menos representada, y los problemas de clasificación excesivamente desequilibrados pueden sesgar la solución obtenida por la máquina de vectores soporte hacia la clase más numerosa [WC03][TZC+09]. En este proyecto se emplea un método de submuestreo aleatorio de la clase mayoritaria, en este caso la clase negativa, teniendo que las dos clases quedan igualmente representadas en el conjunto de entrenamiento T (el número máximo de muestras por clase viene dado por la clase menos numerosa, la clase positiva). Además, de esta forma se reduce el coste computacional en la etapa de entrenamiento de la SVM sin producir una disminución sustancial en sus prestaciones.

Para la SVDD, el conjunto de entrenamiento T está formado únicamente por pacientes de la clase negativa.

La predicción de la mortalidad de los pacientes se obtiene a partir de la salida blanda de las SVMs. A partir de dicha salida, como medida de precisión en la clasificación, se define la precisión promedio de la SVM como:

$$ppSVM = \frac{1}{P}\sum_{p=1}^{p=P}\frac{p}{pos_p} \tag{4.3}$$

donde P es el número de pacientes del conjunto de *test* S de la clase positiva, y pos_p la posición que ocupa un paciente p del conjunto de *test* S de la clase positiva en la salida blanda de la SVM previamente ordenada de forma que los pacientes del conjunto de *test* S de la clase positiva ocupen las primeras posiciones.

Para obtener la precisión promedio de las SVMs previamente se optimizan los parámetros de las SVMs C y σ de forma empírica mediante un proceso de validación cruzada sobre las muestras de entrenamiento para evitar el sobreajuste. El procedimiento seguido para obtener la precisión promedio de las SVMs se describe en el Capítulo 5.

Capítulo 5

5. Experimentos y resultados

En este capítulo se describen los experimentos que se realizaron para comparar el resultado predictivo de los algoritmos propuestos con respecto al resultado obtenido utilizando el índice APACHE II como estimador *baseline*. A continuación se presentan los diferentes resultados experimentales obtenidos que muestran dicha comparación.

5.1. Experimentos

A partir del conjunto de pacientes, dado un índice APACHE II umbral (valor superior de índice APACHE II según los grupos APACHE II definidos en el Capítulo 2), se define la clase positiva, que son los pacientes que mueren en el servicio de la UCI con índice APACHE II menor al índice APACHE II umbral elegido (pacientes etiquetados con +1), y la clase negativa, a la que pertenecen el resto de pacientes que no son de la clase positiva (pacientes etiquetados con -1). En términos de la *one-class* SVC, la clase positiva corresponde con la clase *outlier* y la clase negativa con la clase *target*. En la

Figura 10 se muestra una representación aproximada en 2-D de los pacientes de la UCI según su distancia, etiquetados según los grupos APACHE II definidos en el Capítulo 2.

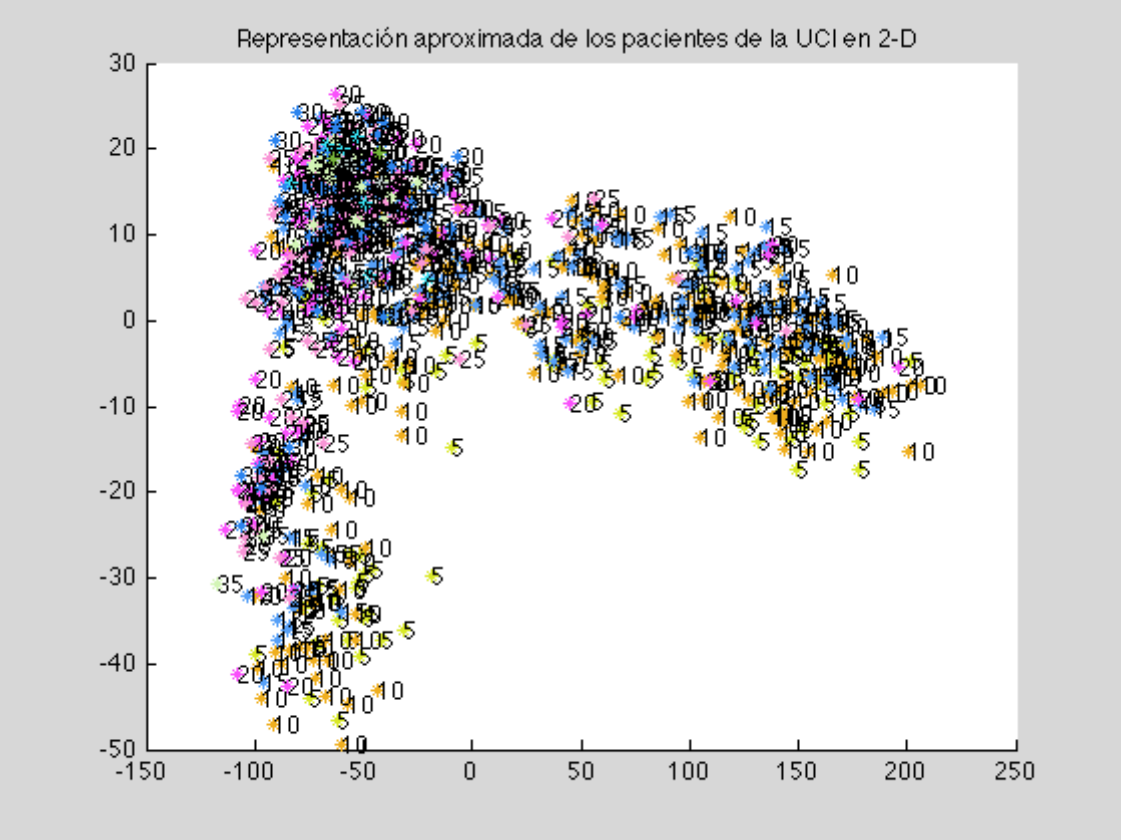

Figura 10. Representación aproximada de los pacientes de la UCI en 2-D

Previamente a la obtención de la precisión promedio de las SVMs se optimizan los parámetros de las SVMs C y σ de forma empírica mediante un proceso de validación cruzada sobre las muestras de entrenamiento para evitar el sobreajuste. Los parámetros de las SVMs C y σ toman el valor de una secuencia creciente exponencialmente, y se realiza una búsqueda en rejilla para obtener el par $\left(C^*,\sigma^*\right)$ con el que se obtiene una mayor precisión promedio. Primero se usa una rejilla gruesa para identificar la mejor zona, para posteriormente usar una búsqueda en rejilla más fina sobre esa zona. Para la LS-SVC $C = \sigma = 1e-5,1e-4,...,1e+5$, y para la SVDD el valor de C está acotado al

valor $C \geq 1/n$, siendo n el número de muestras para el entrenamiento, con $N = 830$ el número total de pacientes considerados en el conjunto de datos de entrada X, por lo tanto, en este caso C toma los valores $C = 1e-2, 1e-14, ..., 1e+8$, y σ toma los mismos valores que en el caso de la LS-SVC. En la *Figura 11* y *Figura 12* se muestra un ejemplo de búsqueda en rejilla para obtener los parámetros óptimos de las SVMs $\left(C^*, \sigma^*\right)$.

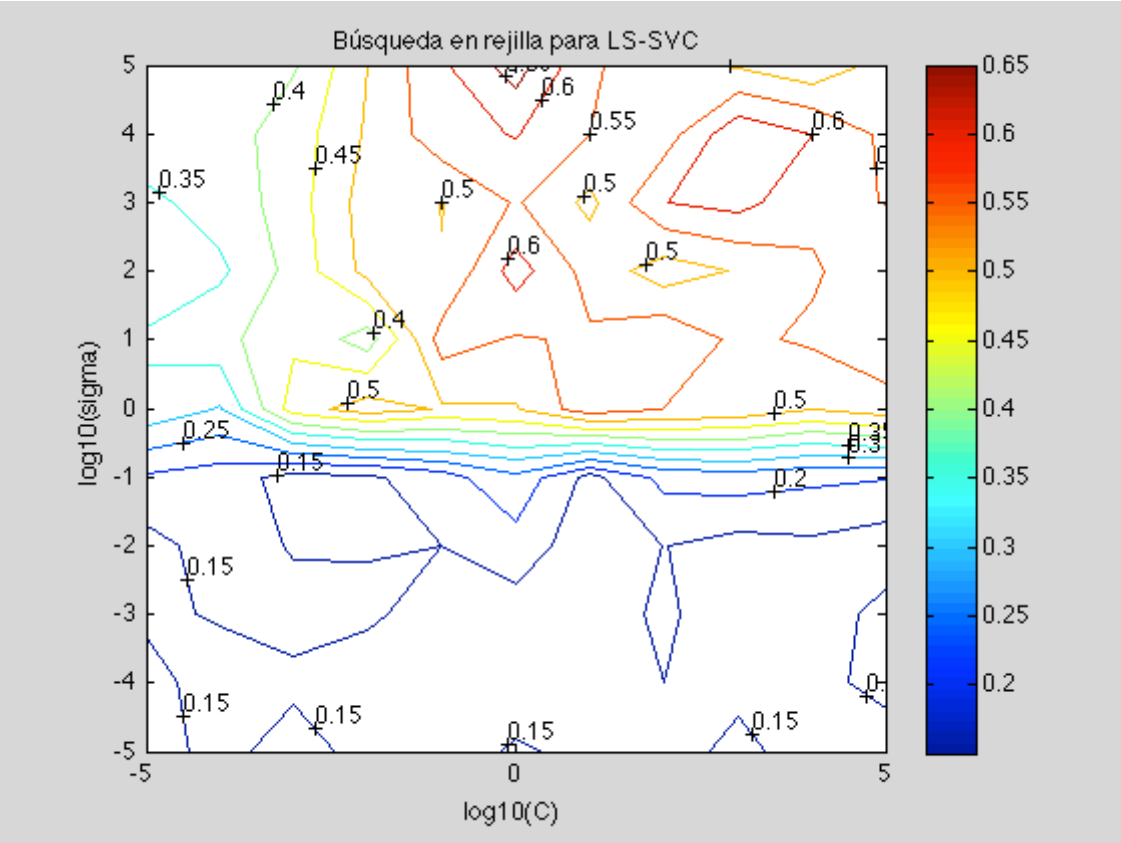

Figura 11. Búsqueda en rejilla de parámetros (C,σ) para la LS-SVC

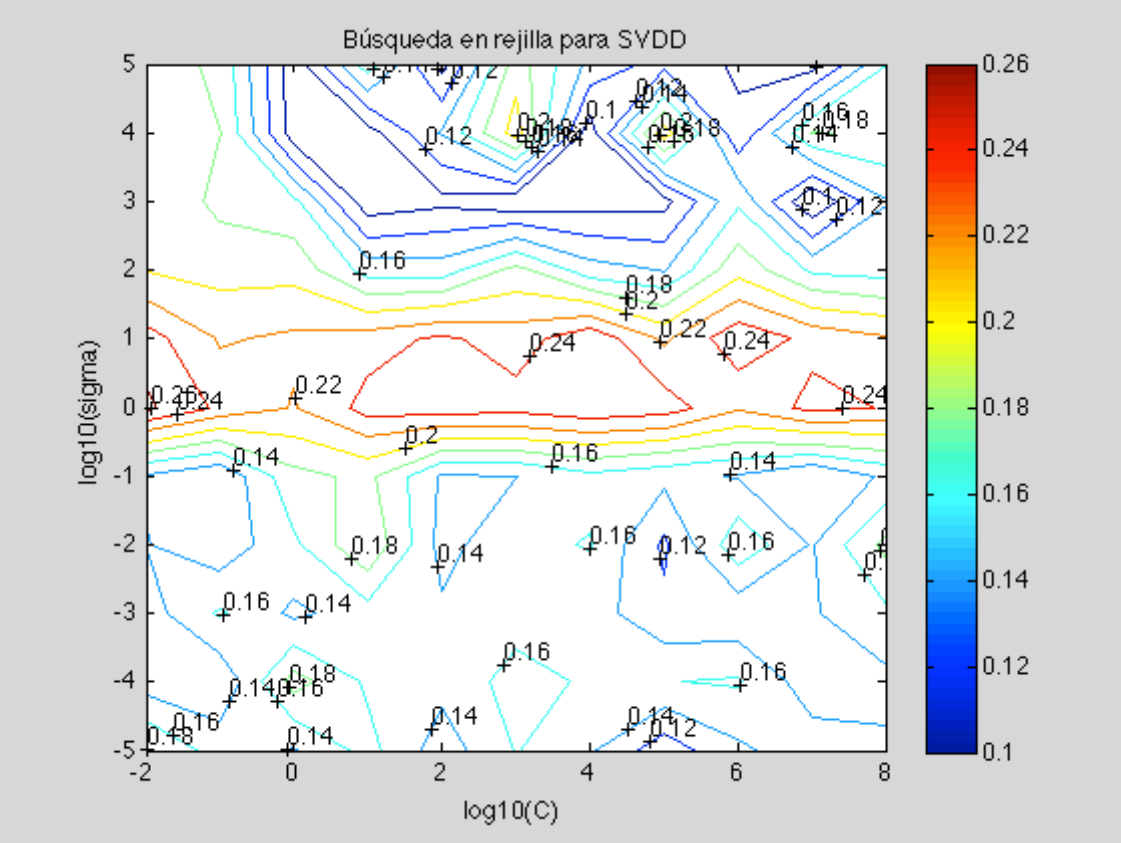

Figura 12. Búsqueda en rejilla de parámetros (C,σ) para la SVDD

Para obtener la precisión promedio de las SVMs se realiza un proceso de validación cruzada anidada, donde la validación cruzada interna se utiliza para elegir los parámetros óptimos de las SVMs $\left(C^*,\sigma^*\right)$, y la externa para estimar la precisión promedio de las SVMs. El procedimiento seguido para obtener la precisión promedio de las SVMs se describe a continuación:

1. El conjunto de pacientes X etiquetado, se divide en 5 grupos mediante validación cruzada *5-fold* (cada uno de los grupos está formado por el 20% de las muestras del conjunto de pacientes X).

2. El conjunto de *test* S_k, k=1,...,5 está formado por uno de los grupos creados anteriormente, y el resto de los grupos forman el conjunto de entrenamiento T_k (que corresponde con el 80% de las muestras del conjunto de pacientes X).
3. El conjunto de entrenamiento T_k se divide, a su vez, en 5 subgrupos mediante validación cruzada *5-fold*, donde uno de los subgrupos corresponde con el subconjunto de *test* $S_k(T_k) \subset T_k$ (que corresponde con el 16% de las muestras del conjunto de pacientes X), y el resto los subgrupos forman el conjunto de entrenamiento $T_k(T_k) \subset T_k$ (que corresponde con el 64% de las muestras del conjunto de pacientes X).
4. Para cada par (C,σ) se entrenan las SVMs con los subconjuntos de entrenamiento $T_k(T_k)$, obteniendo un valor medio de precisión promedio para cada una de las SVMs, mediante la evaluación de los correspondientes subconjuntos de *test* $S_k(T_k)$.
5. Con el par (C^*,σ^*) que se obtiene el máximo valor medio de precisión promedio, se entrenan las SVMs con los conjuntos de entrenamiento T_k, obteniendo un valor medio de precisión promedio final para cada una de las SVMs, mediante la evaluación de los correspondientes conjuntos de *test* S_k.

El valor medio de la precisión promedio final de las SVMs se compara con el mayor valor medio de precisión promedio final que se puede obtener si se utiliza como salida blanda del clasificador el índice APACHE II (estimador *baseline*). En este caso, el mayor valor medio de precisión promedio final se obtiene dando preferencia en orden para las primeras posiciones a los pacientes de la clase positiva frente a los pacientes de la clase negativa con el mismo índice APACHE II.

Para la implementación de las SVMs se hace uso de la librería de Matlab *bioinformatics toolbox* (de la versión Matlab 7.8 R2009a).

5.2. Resultados

Los resultados de valor medio de precisión promedio final se obtienen utilizando dos modelos diferentes, como se describe a continuación:

1. Modelo utilizando la información disponible en el momento del ingreso de los pacientes en el servicio de la UCI. En este caso, para la caracterización de los pacientes solo se utilizan las variables *Edad*, *Sexo*, *Días transcurridos hasta el ingreso en UCI*, *Procedencia*, *Causa de ingreso al hospital*, *Enfermedad*, *APACHE II*, *Antecedentes*, y *Diagnósticos*.
2. Modelo en el que se incluyen las variables que contienen la información de la evolución de los pacientes en el servicio de la UCI. Además de las variables de entrada en el servicio de la UCI, según la evolución del paciente se conocen las variables *Técnicas*, *Complicaciones*, y *Días en el servicio de la UCI.*

Para ambos modelos, se presenta una comparación del valor medio de precisión promedio final obtenido por los sistemas LS-SVC y SVDD propuestos, respecto al máximo valor medio de precisión final que se obtiene si se utiliza como salida blanda del clasificador el índice APACHE II (sistema de referencia).

5.2.1. Resultados obtenidos utilizando el modelo en el que solo se usan las variables de entrada en el servicio de la UCI

En la *Tabla 13* se muestran los resultados obtenidos, en función del índice APACHE II umbral, del valor medio de precisión promedio final junto con el error estándar cometido en la medida de dicho valor (standard deviation, STD), los cuales se representan gráficamente en la *Figura 13*.

Se puede observar que, para un índice APACHE II umbral de 5, el valor medio de precisión promedio final es cero, debido a que no existen pacientes con un índice APACHE II menor de 5 que mueran en el servicio de la UCI (no se puede definir la clase positiva). Para los valores de índice APACHE II umbral 10 y 15, con la LS-SVC y SVDD se tiene un error estándar de precisión promedio final significativo, debido a que existen pocos pacientes con índices APACHE II menores de 10 y 15 que mueran en el servicio de la UCI, y por lo tanto, no es preciso el resultado del valor medio de precisión promedio final obtenido para esos índices de APACHE II umbral con la LS-SVC y SVDD. A partir del índice APACHE II umbral 15, con la LS-SVC se obtiene el mayor valor medio de precisión promedio final hasta el índice APACHE II umbral 40, que es superado por el sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II. Los peores resultados de valor medio de precisión promedio final se obtienen con la SVDD, excepto para el índice APACHE II umbral 20, cuyo resultado es comparable con el obtenido por el sistema de referencia.

Precisión promedio usando solo las variables de entrada en el servicio de la UCI											
		Índice APACHE II umbral									
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
ppSVM LS-SVC	**Media**	0.0000	0.1338	0.0598	0.2695	0.3683	0.4339	0.5402	0.5580	0.5616	0.5715
	STD	0.0000	0.1253	0.0249	0.0392	0.0550	0.0381	0.1667	0.0878	0.0645	0.0582
ppSVM SVDD	**Media**	0.0000	0.0334	0.1342	0.1178	0.1992	0.2278	0.3278	0.3232	0.3500	0.3607
	STD	0.0000	0.0403	0.2291	0.0481	0.0220	0.0676	0.0623	0.0576	0.0739	0.0386
ppSVM APACHEII	**Media**	0.0000	0.0104	0.0366	0.1072	0.2511	0.3817	0.5103	0.5626	0.6251	0.6650
	STD	0.0000	0.0017	0.0064	0.0125	0.0151	0.0398	0.0501	0.0762	0.0595	0.0355

Tabla 13. Precisión promedio usando solo las variables de entrada en el servicio de la UCI

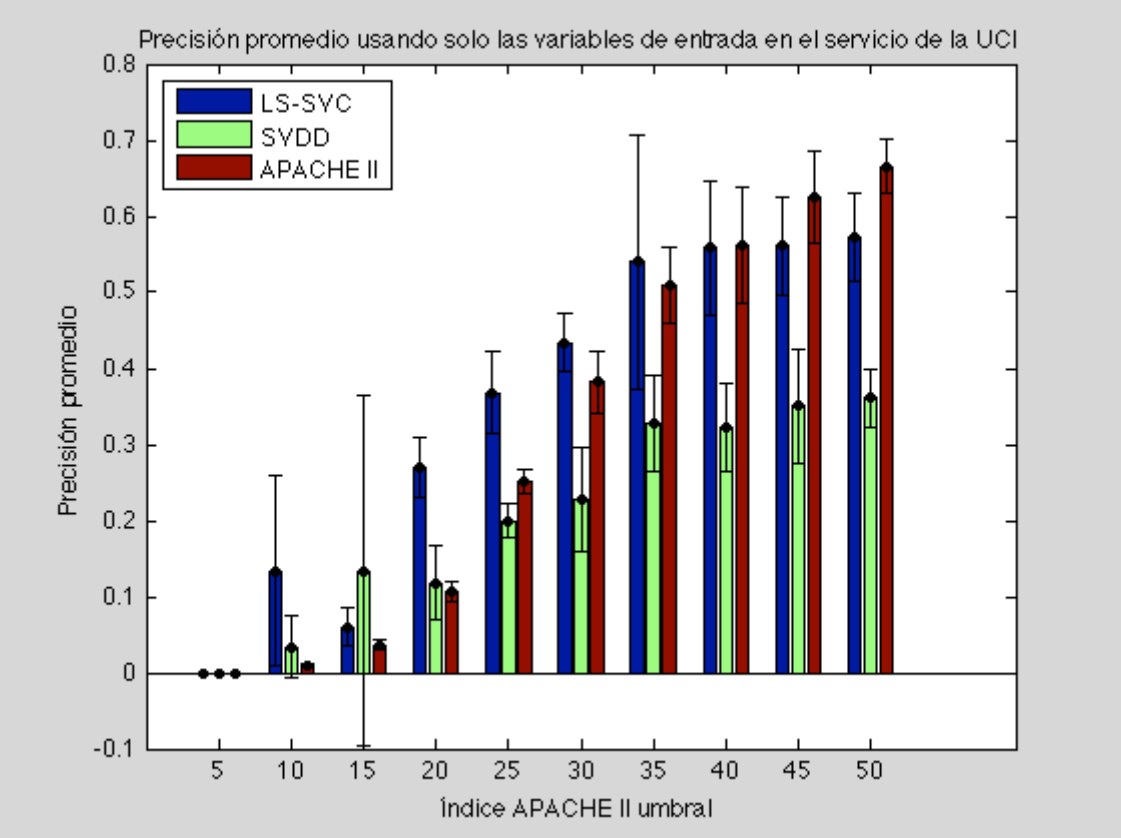

Figura 13. Precisión promedio usando solo las variables de entrada en el servicio de la UCI

5.2.2. Resultados obtenidos utilizando el modelo en el que se usan todas las variables en el servicio de la UCI

En la *Tabla 14* se muestran los resultados obtenidos, en función del índice APACHE II umbral, del valor medio de precisión promedio final junto con el error estándar cometido en la medida de dicho valor (standard deviation, STD), los cuales se representan gráficamente en la *Figura 14*.

Se puede observar que, para un índice APACHE II umbral de 5, el valor medio de precisión promedio final es cero, debido a que no existen pacientes con un índice

APACHE II menor de 5 que mueran en el servicio de la UCI (no se puede definir la clase positiva). Para el valor de índice APACHE II umbral 10, con la LS-SVC y SVDD se tiene un error estándar de precisión promedio final significativo, debido a que existen pocos pacientes con índice APACHE II menor de 10 que mueran en el servicio de la UCI, y por lo tanto, no es preciso el resultado del valor medio de precisión promedio final obtenido para dicho índice de APACHE II umbral con la LS-SVC y SVDD. Con la LS-SVC se obtiene el mayor valor medio de precisión promedio final, valor que no es superado por ninguno de los otros dos sistemas utilizados. Los peores resultados de valor medio de precisión promedio final se obtienen con la SVDD, excepto para los índices de APACHE II umbral 15 y 20, cuyo resultado es comparable con el obtenido por el sistema de referencia.

Precisión promedio usando todas las variables en el servicio de la UCI											
		Índice APACHE II umbral									
		5	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
ppSVM LS-SVC	**Media**	0.0000	0.5074	0.3220	0.3923	0.5523	0.6907	0.7942	0.7926	0.8432	0.8509
	STD	0.0000	0.4907	0.0937	0.0853	0.1384	0.0461	0.0465	0.0557	0.0576	0.0299
ppSVM SVDD	**Media**	0.0000	0.2104	0.0510	0.1173	0.2190	0.3099	0.3657	0.3752	0.3959	0.4301
	STD	0.0000	0.4414	0.0203	0.0295	0.0262	0.0618	0.0834	0.1177	0.0934	0.1269
ppSVM APACHEII	**Media**	0.0000	0.0102	0.0369	0.1060	0.2521	0.3823	0.5039	0.5713	0.6417	0.6721
	STD	0.0000	0.0011	0.0067	0.0093	0.0385	0.0622	0.0435	0.0781	0.1412	0.1270

Tabla 14. Precisión promedio usando todas las variables en el servicio de la UCI

Con el modelo en el que se usan todas las variables en el servicio de la UCI, con la LS-SVC y SVDD se obtienen mejores resultados de valor medio de precisión promedio final, mejoras, en media, del 39.16% y del 16.17% respectivamente, en comparación con los resultados obtenidos con el modelo en el que solo se usan las variables de entrada en el servicio de la UCI.

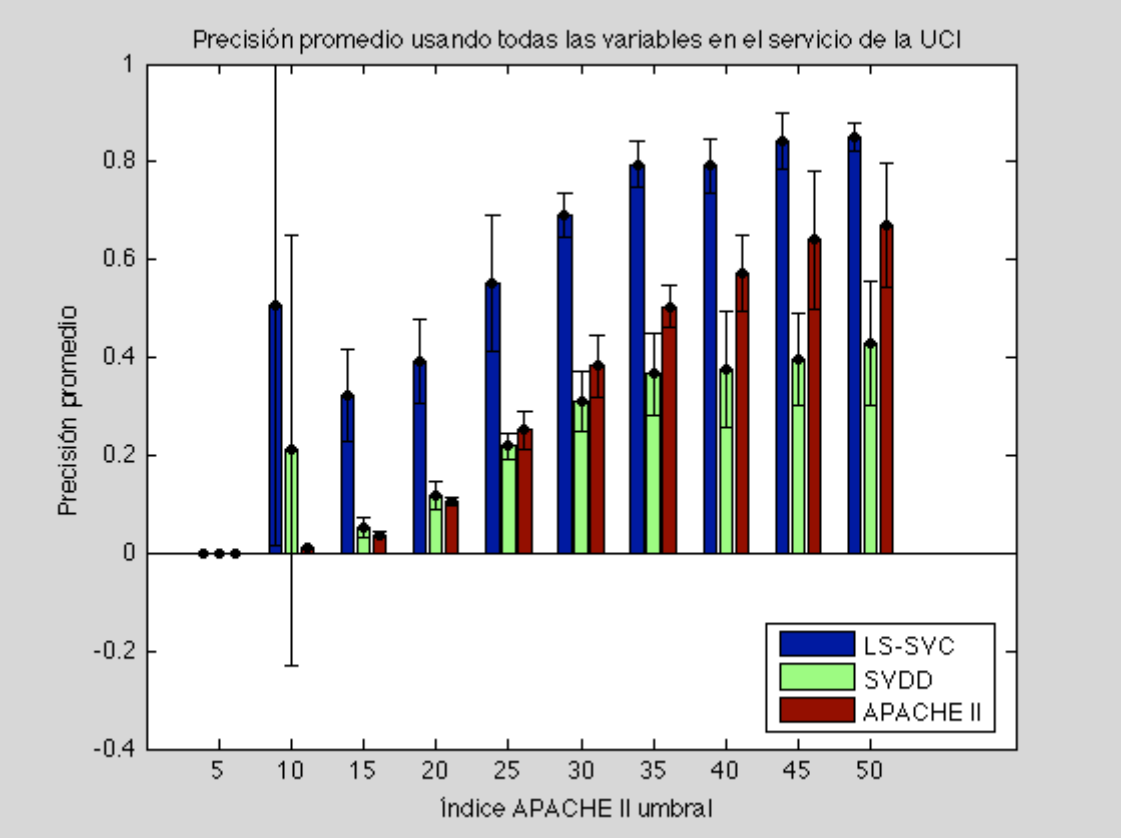

Figura 14. Precisión promedio usando todas las variables en el servicio de la UCI

Además, se puede observar en los resultados obtenidos con ambos modelos, que el valor medio de precisión promedio final es mayor cuanto más grande es el índice APACHE II umbral, es decir, si se conocen más ejemplos de pacientes que mueren en el servicio de la UCI. Por lo tanto, como era de esperar en un principio, cuanto más información se tenga respecto a los pacientes, y mayor sea el número de ejemplos de pacientes que mueren en el servicio de la UCI, mejores serán los resultados obtenidos con los sistemas propuestos respecto al resultado obtenido con el sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II.

Otra posible vía de investigación consistiría en la modificación de las funciones *kernel* empleadas, de modo que se primen las características más relevantes de los pacientes de la UCI.

Si se comparan los dos sistemas propuestos, se puede observar que con la LS-SVC, en ambos modelos, se obtienen mejores resultados de valor medio de precisión promedio final que con la SVDD, presentando con la LS-SVC, además, una complejidad computacional mucho menor que con la SVDD. Utilizando el modelo en el que solo se usan las variables de entrada en el servicio de la UCI, con la LS-SVC se obtiene una mejora, en media, en el resultado del valor medio de precisión promedio final del 56.95% respecto al resultado obtenido con la SVDD. En el modelo en el que se usan todas las variables en el servicio de la UCI, con la LS-SVC se obtiene una mejora, media, en el resultado del valor medio de precisión promedio final del 40.68% respecto al resultado obtenido con la SVDD.

Finalmente, se discuten los resultados obtenidos por el sistema LS-SVC respecto al máximo valor medio de precisión final que se obtiene si se utiliza como salida blanda del clasificador el índice APACHE II (sistema de referencia). Utilizando el modelo en el que solo se usan las variables de entrada en el servicio de la UCI, con la LS-SVC se obtiene una mejora (no estadísticamente significativa), en media, en el resultado del valor medio de precisión promedio final del 9,9%, respecto al resultado obtenido con el sistema de referencia, y en el modelo en el que se usan todas las variables en el servicio de la UCI, con la LS-SVC se obtiene una mejora (estadísticamente significativa), media, en el resultado del valor medio de precisión promedio final del 44.71% respecto al resultado obtenido con el sistema de referencia.

Capítulo 6

6. Conclusiones y líneas futuras de trabajo

En este capítulo se resume el estudio realizado en este proyecto, analizando los resultados obtenidos y el cumplimiento de los objetivos planteados en el Capítulo 1. Así mismo, se proponen las líneas futuras de trabajo que se consideran de especial interés.

6.1. Conclusiones

En este proyecto se realiza un estudio experimental relativo a la aplicación de la SVM para la predicción de la mortalidad de los pacientes en el servicio de la UCI, con el objetivo de estudiar si existe o no una clase definida para los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo, determinando si los resultados obtenidos del estudio son significativos en el sentido de

ayuda a la toma de decisiones respecto a los pacientes que mueren en el servicio de la UCI.

En todas las terapias intensivas se utiliza este índice como marcador pronóstico al ingreso de los pacientes críticos, lo que permite estratificar la complejidad de los pacientes internados, observando, en base a los datos obtenidos, que existe una relación directamente proporcional entre el índice APACHE II y la mortalidad. Sin embargo, las circunstancias particulares de cada hospital (características demográficas de los pacientes, capacitación del personal, etc.) hacen que las predicciones efectuadas por el índice APACHE II no siempre se cumplan. Por otra parte, cada vez son más las patologías donde este índice es un marcador independiente de la mortalidad, como es en el caso de las pancreatitis. El punto de corte del índice APACHE II que predice la mortalidad es de 26, lo que significa que aquellos pacientes con índice APACHE II mayor de 26 tendrán más probabilidades de morir. Si bien, la literatura publicada demuestra que los puntos de corte se encuentran en distintos valores de índice APACHE II al observado en este caso, debido a que responde exclusivamente a los datos de este proyecto.

Los pacientes están caracterizados por variables tanto binarias como cualitativas y cuantitativas, representando los datos disponibles en el momento del ingreso de los pacientes, más su evolución, en el servicio de la UCI. Como resultado del estudio estadístico de las variables que caracterizan a los pacientes, se han obtenido diferentes indicadores, mostrados en el Capítulo 4, que pueden ser de interés para su utilización en la UCI del Hospital Clínico Universitario de Valladolid.

La máquina de vectores soporte, en su modalidad para clasificación (SVC), se utiliza para predecir a posteriori la mortalidad de cada uno de los pacientes de la UCI, es decir, saber si un paciente tiene una evolución favorable en el servicio de la UCI o, por el contrario, muere, comparando los resultados obtenidos mediante LS-SVC y SVDD, respecto al resultado que se puede obtener utilizando el índice APACHE II como estimador *baseline* (sistema de referencia). La máquina de vectores soporte tiene un modelo estado del arte no superado hasta el momento, SVC, que le confieren a priori ciertas ventajas respecto a

otras técnicas empleadas, obteniendo, una vez que se han fijado adecuadamente los parámetros de la SVM, excelentes precisiones y buenas propiedades de generalización. Los resultados de clasificación de las SVMs se han obtenido utilizando dos modelos diferentes, como se describe a continuación:

1. Modelo utilizando la información disponible en el momento del ingreso de los pacientes en el servicio de la UCI.
2. Modelo en el que se incluyen las variables que contienen la información de la evolución de los pacientes en el servicio de la UCI.

Para ambos modelos, se presenta una comparación de la clasificación obtenida por los sistemas LS-SVC y SVDD propuestos, respecto al sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II, en el que se muestra, que con el sistema LS-SVC propuesto se consiguen resultados competitivos respecto al resultado obtenido con el sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II, tanto en el modelo en el que solo se usan las variables de entrada en el servicio de la UCI, obteniendo una mejora, en media, del 9,9% (no estadísticamente significativa), como en el modelo en el que se usan todas las variables en el servicio de la UCI, obteniendo una mejora, en media, del 44.71% (estadísticamente significativa), y **se puede afirmar, según los resultados obtenidos, que sí existe una clase definida para los pacientes que mueren en la UCI teniendo un factor predictivo de mortalidad (dado por el índice APACHE II) bajo**, siendo necesario un mayor esfuerzo investigador para determinar si los resultados obtenidos son concluyentes en el sentido de ayuda a la toma de decisiones en la gestión de la UCI.

Por lo tanto, resulta necesario considerar una serie de líneas futuras de investigación que son de especial interés.

6.2. Líneas futuras de trabajo

En esta sección se proponen, después de la evaluación experimental del estudio realizado, una serie de líneas futuras de investigación que son de especial interés:

- Considerar más variables de estudio que caractericen a los pacientes y tener una base de datos de pacientes de la UCI más grande en la que exista un mayor número de ejemplos de pacientes que mueren en el servicio de la UCI, ya que se ha observado que cuanto más información se tenga respecto a los pacientes, y mayor sea el número de ejemplos de pacientes que mueren en el servicio de la UCI, mejores serán los resultados obtenidos con los sistemas propuestos respecto al resultado obtenido con el sistema de referencia basado en utilizar como salida blanda del clasificador el índice APACHE II.
- Tratamiento de bases de datos de pacientes de la UCI de diferentes hospitales, con el objetivo de determinar si los resultados obtenidos se pueden generalizar.
- Estudio de la posibilidad de modificar las funciones de *kernel* empleadas en las SVMs de modo que primen las características más relevantes de los pacientes de la UCI.
- Determinar, si es posible, las causas por las que mueren los pacientes de la UCI que tienen un factor predictivo de mortalidad (dado por el índice APACHE II) bajo.

Capítulo 7

7. Planificación y presupuesto

En este último capítulo de la memoria se incluye la planificación detallada del proyecto, así como los costes, según la duración y los recursos de la planificación, que ha supuesto el desarrollo del proyecto.

7.1. Planificación

Para realizar la planificación se ha utilizado la herramienta GanttProject, como software libre y de código abierto, *opensource*, de planificación y gestión de proyectos [GANTTPROJECT].

Primero se muestra una descripción general de las fases principales del proyecto, y a continuación, se estructura la planificación en tareas mediante un diagrama de Gantt.

7.1.1. Fases del proyecto

El proyecto se ha realizado en varias fases, realizando algunas de ella en paralelo. A continuación, se enumeran las diferentes fases del proyecto con una breve descripción del propósito de cada una de ellas:

1. **Fase 1. Planteamiento**. Se plantea la propuesta inicial, realizando un estudio previo sobre los objetivos del proyecto y se analizan las posibles técnicas a utilizar.
2. **Fase 2. Tratamiento de la Base de Datos**. La Base de Datos (B. D.), proporcionada por el Hospital Clínico Universitario de Valladolid, se importa a Matlab para posteriormente trabajar mejor la parte de algoritmia. Existió un compromiso de colaboración y apoyo de los investigadores del Hospital Clínico Universitario de Valladolid para entender e interpretar los diferentes registros de la base de datos. Se codifican y estudian las variables seleccionadas, obteniendo como resultado diferentes indicadores de interés.
3. **Fase 3. Aplicación de la SVM**. Se implementan las SVMs, haciendo uso de la librería de Matlab *bioinformatics toolbox*, en base al estudio realizado en la fase previa. Durante las simulaciones realizadas se corrigen los diferentes errores en la implementación y se optimizan los parámetros de las SVMs para obtener los resultados finales de clasificación.
4. **Fase 4. Presentación**. Se edita la memoria del proyecto y se prepara la presentación, para finalmente realizar la lectura y defensa del proyecto.

7.1.2. Diagrama de Gantt

En la *Figura 15* se muestran, mediante un diagrama de Gantt, las tareas que se han realizado en este proyecto, agrupadas en las diferentes fases que se han definido anteriormente.

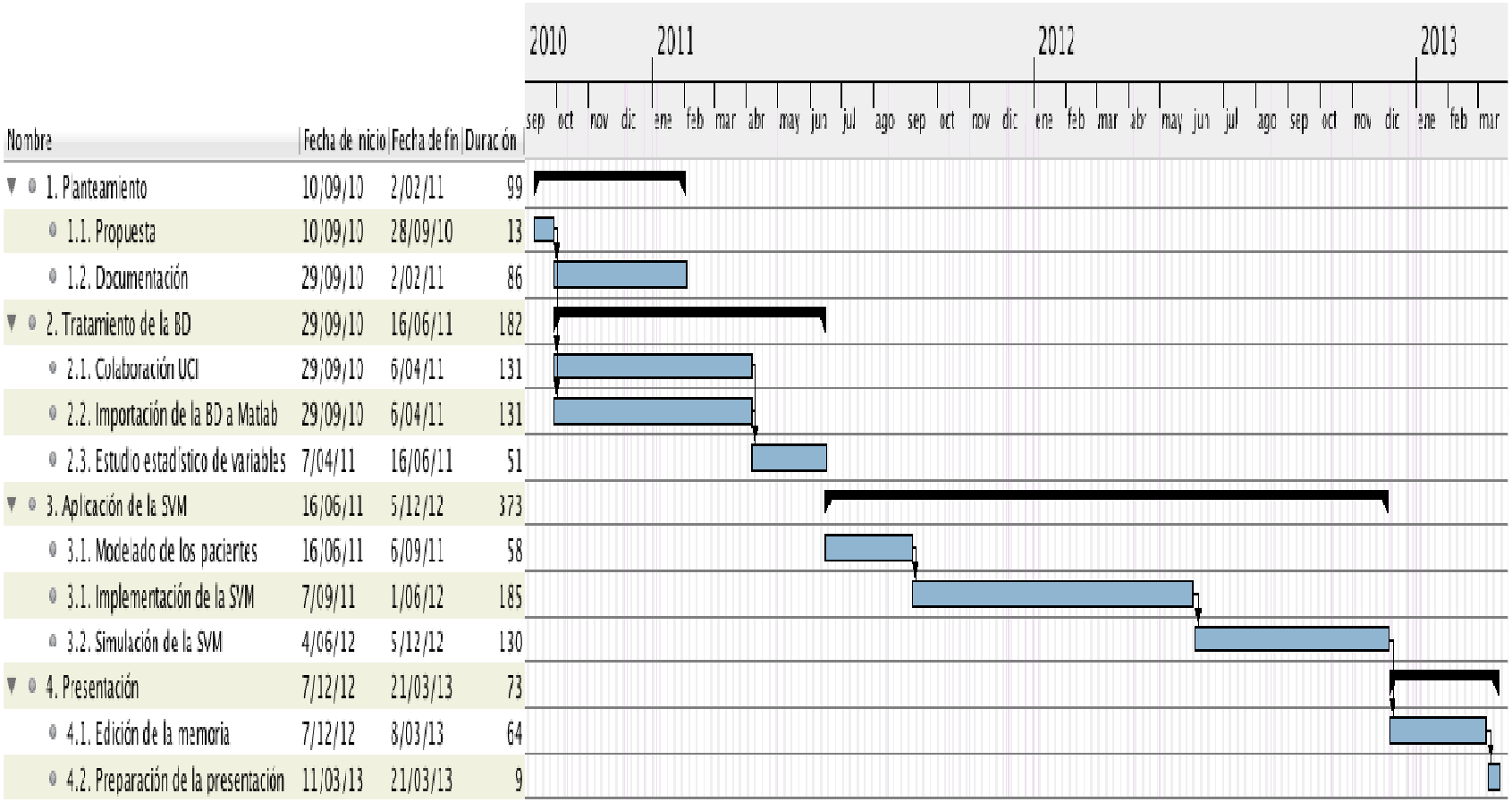

Figura 15. Diagrama de Gantt con la planificación del proyecto

7.2. Presupuesto

En esta sección se desglosa, de forma detallada, el presupuesto del proyecto, especificando los diferentes costes que han sido necesarios para su realización.

7.2.1. Número de horas dedicadas al proyecto

A partir de la planificación realizada se calcula el número de horas dedicadas en total al proyecto como se muestra en la *Tabla 15*.

Fase	Duración (días)	Horas/día	Horas dedicadas
1. Planteamiento	99	2	198
2. Tratamiento de la Base de Datos	182	2	364
3. Aplicación de la SVM	373	1	373
4. Presentación	73	4	292
		Total	**1227**

Tabla 15. Número de horas dedicadas al proyecto

Por lo tanto, el número de horas dedicadas en total al proyecto es la suma de las horas dedicadas a cada una de las fases. El coste en horas de la totalidad del proyecto es de 1227 horas.

La duración del proyecto ha sido de aproximadamente 31 meses. El proyecto se ha alargado tanto en el tiempo debido a que se ha tenido que compatibilizar la realización del proyecto con el trabajo.

7.2.2. Costes de personal

En la realización de este proyecto ha intervenido un Ingeniero *Senior* de Telecomunicación y un Ingeniero de Telecomunicación. El coste de personal se muestra en la *Tabla 16*.

Categoría	Dedicación (hombre/mes) [a)]	Coste hombre/mes (Euros)	Coste (Euros)
Ingeniero *Senior*	0,47	4.289,54	2.016,08
Ingeniero	9,34	2.694,39	25.165,60
		Total	**27.181,69**

Tabla 16. Costes de personal

[a)] 1 hombre/mes = 131,25 horas. Máximo anual de dedicación de 12 hombres/mes (1575 horas). Máximo anual para PDI de la Universidad Carlos III de Madrid de 8,8 hombres/mes (1.155 horas).

7.2.3. Costes de equipos

En la *Tabla 17* se consideran los costes de los equipos que han sido necesarios para la realización del proyecto.

Descripción	Coste (Euros)	%Uso dedicado al proyecto	Dedicación (meses)	Periodo de depreciación	Coste imputable [b)]
Portátil MacBook Pro 13”	1.100,00	100	31	60	568,33
Memoria USB 8 GB de Kingston	20,00	100	31	60	10,33
				Total	**578,67**

Tabla 17. Costes de equipos

[b)] Siendo la fórmula del cálculo de la amortización:

$$\frac{A}{B} \times C \times D, \quad \begin{array}{l} A = \text{número de meses desde la fecha de facturación en que el equipo es utilizado} \\ B = \text{periodo de depreciación (60 meses)} \\ C = \text{coste del equipo} \\ D = \% \text{ de uso que se dedica al proyecto} \end{array} \tag{7.1}$$

7.2.4. Costes de software y licencias

En la *Tabla 18* se muestran los costes de las herramientas software que han sido necesarias para la realización del proyecto.

Descripción	Unidades	Coste unidad (Euros)	Coste (Euros)
Microsoft Office 2011	1	110,00	110,00
Matlab	1	300,00	300,00
GanttProject	1	0	0
		Total	**410,00**

Tabla 18. Costes de software y licencias

7.2.5. Resumen de costes

Para el cálculo del presupuesto total, en la *Tabla 19* se resume el presupuesto de los costes totales. En la realización de este proyecto no se tienen costes por subcontratación de tareas, ni costes de funcionamiento (fungible, viajes y dietas, otros). A la suma del presupuesto de los costes totales se añade un 20% en concepto de costes indirectos, lo que hace introducir otros costes que no se han tenido en cuenta al realizar el presupuesto

y equilibra el riesgo del proyecto. Además, al presupuesto total del proyecto se le adiciona el 21% de IVA.

Descripción	Presupuesto Costes Totales (Euros)
Personal	27.181,69
Amortización	578,67
Software y licencias	410,00
Subcontratación de tareas	0,00
Costes de funcionamiento	0,00
Costes indirectos	5.634,07
TOTAL	**33.804,43**

IVA 21%	7098,93
TOTAL (con IVA 21%)	**40.903,36**

Tabla 19. Resumen de costes del presupuesto total del proyecto

El presupuesto total de este proyecto asciende a la cantidad de CUARENTA MIL NOVECIENTOS TRES EUROS CON TREINTA Y SEIS CENTIMOS.

Leganés, a 8 de marzo de 2013

El Ingeniero proyectista

Fdo.: David Toledo Navarro

Bibliografía

[ABR64] Aizerman, M. A., Braverman, E. M. y Rozonoer, L. I.: *Theoretical Foundations of the Potential Function Method in Pattern Recognition Learning*. Automation and Remote Control, 25, pp. 821-837, 1964

[AMV08] Assareh, A., Moradi, M.H. y Volkert, L. G.: *A Hybrid Random Subspace Classifier Fusion Approach for Protein Mass Spectra Classification*, en Marchiori, E. y Moore, J. H. (Ed.): *EvoBIO 2008*, LNCS, vol. 4973, pp. 1-11. Springer, Heidelberg, 2008

[Bis06] Bishop, C. M.: *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, August 2006

[BGV92] Boser, B. E., Guyon, I. M. y Vapnik, V. N.: *A Training Algorithm for Optimal Margin Classifiers*, en: *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, pp. 144-152. ACM Press, Nueva York, Nueva York, Estados Unidos, 1992

[Bur98] Burges, C. J. C.: *A Tutorial on Support Vector Machines for Pattern Recognition*. Data Mining and Knowledge Discovery, 2(2), pp. 121-167, 1998

[CGM99] Costa, J. I., Gomes do Amaral, J. L. y Munechika, M.: *Severity and Prognosis in Intensive Care: Prospective Application of the Apache II Index*. Rev Paul Med, 117(5):205-14, 1999

[Che10] Chen, J.: *Biological Data Mining*, Chapman & Hall/CRC, 2010

[CLJ89] Chang, R. W., Lee, B. y Jacobs, S.: *Accuracy of Decisions to Withdraw Therapy in Critically Ill Patients: Clinical Judgment Versus a Computer Model*. Crit Care Med, 17:1091-7, 1989

[Coi97] Coiera, E.: *Guide to Medical Informatics, the Internet and Telemedicine*, Chapman & Hall, 1997

[Cov65] Cover, T. M.: *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition*. IEEE Transactions on Electronic Computers, 14, pp. 326-334, 1965

[CV95] Cortes, C. y Vapnik, V. N.: *Support-Vector Networks*. Machine Learning, 20(3), pp. 273-297, 1995

[DBK+97] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J. y Vapnik, V. N.: *Support Vector Regression Machines*, en Mozer, M., Jordan, M. y Petsche, T. (Ed.): *Advances in Neural Information Processing Systems 9*, pp. 155-161. MIT Press, Cambridge, Massachusetts, Estados Unidos, 1997

[GANTTPROJECT] *GPL-licensed Project Management Software*. Disponible [Internet]: http://www.ganttproject.biz/ [12 de febrero de 2013]

[Gar09] García-Gómez, J. M.: *Pattern Recognition Approaches for Biomedical Data in Computer-Assistedcancer Research*. Tesis Doctoral, Universidad Politécnica de Valencia, Valencia, España, 2009

[GBV93] Guyon, I. M., Boser, B. E. y Vapnik, V. N.: *Automatic Capacity Tuning of Very Large VC-Dimension Classifiers*, en Hanson, S. J., Cowan, J. D. y Giles, C. L. (Ed.): *Advances in Neural Information Processing Systems 5*, pp. 147-155. Morgan Kaufmann Publishers, San Francisco, California, Estados Unidos, 1993

[Gir99] Giráldez, J. I.: *Modelo de Toma de Decisiones y Aprendizaje en Sistemas Multi-Agente*. Tesis Doctoral, Facultad de Informática, Dpto. Inteligencia Artificial, Universidad Politécnica de Madrid, Madrid, España, 1999

[Gow66] Gower, J. C.: Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis. Biometrika, 53, 325-338, 1966

[Gow71] Gower J.C.: *A General Coefficient of Similarity and Some of its Properties*. Biometrics, Vol. 27, pp. 857-872, 1971

[GR99] Gunning, K. y Rowan, K.: *ABC of Intensive Care: Outcome Data and Scoring Systems*. British Medical Journal;319:241-244, 1999

[HHH+98] Hunt, D. L., Haynes, R. B., Hanna, S. E. y Smith, K.: *Effects of Computer Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes: A Systematic Review*. JAMA, 280(15):1339-1346, 1998

[HL89] Hosmer, D. W. y Lemeshow, S.: *Applied Logistic Regression*. New York, John Wiley, 370p, 1989

[KDW+85] Knauss, W. A., Draper, E. A., Wagner, D. P. y Zimmerman, J. E.: *APACHE II: A Severity of Disease Classification System*. Crit Care Med, 13:818, 1985

[KT51] Kuhn, H. W. y Tucker, A. W.: *Nonlinear Programming*, en: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 481-492. Berkeley, California, Estados Unidos, 1951

[KW71] Kimeldorf, G. y Wahba, G.: *Some Results on Tchebycheffian Spline Function*. Journal of Mathematical Analysis and Applications, 33(1), pp. 82-95, 1971

[Lav99] Lavrac, N.: *Selected Techniques for Data Mining in Medicine*. Artificial Intelligence in Medicine 16, 3-23, 1999

[Lis02] Lisboa, P. J. G.: *A Review of Evidence of Health Benefit from Artificial Neural Networks in Medical Intervention*. Neural Networks, 15(1):11-39, 2002

[LFT00] Laurent, G., Furner, O. y Tamatsu, S.: *Effect of Varying the Case Mix on the Standardized Mortality Radio and W Statistic*. Chest;117:1112-1117, 2000

[LRD04] Lesage, A., Ramakers, M. y Daubin, C.: *Complicated Acute Myocardial Infarction Requiring Mechanical Ventilation in the Intensive Care Unit: Prognostic Factors of Clinical Outcome in a Series of 157 Patients*. Crit Care Med, 32(1):100-105, 2004

[Mar08] Márquez, M.: *Clasificación de Sedimentos Clásticos Mediante Máquinas de Vectores Soporte*. Tesis Doctoral, Universidad de los Andes Mérida, Venezuela, 2008

[MATHWORKS] *MATLAB and Simulink for Technical Computing*. Disponible [Internet]: http://www.mathworks.com/ [08 de enero de 2011]

[Mer09] Mercer, J.: *Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations*. Philosophical Transactions of the Royal Society, A 209, pp. 415-446, 1909

[NPA+01] Navia-Vázquez, Á., Pérez-Cruz, F., Artés-Rodríguez, A. y Figueiras-Vidal, A. R.: *Weighted Least Squares Training of Support Vector Classifiers Leading to Compact and Adaptive Schemes*. IEEE Transactions on Neural Networks, 12(5), pp. 1047-1059, 2001

[NW99] Nocedal, J. y Wright, S. J.: *Numerical Optimization*. Springer, Nueva York, Nueva York, Estados Unidos, 1999

[PBA05] Pérez-Cruz, F., Bousoño-Calzón, C. y Artés-Rodríguez, A.: *Convergence of the IRWLS Procedure to the Support Vector Machine Solution*. Neural Computation, 17(1), pp. 7-18, 2005

[PD11] Palma, J. T. y Díez, F. J.: *Análisis de Datos y Toma de Decisiones en Medicina a través de la Inteligencia Artificial*. Guadalajara, España, 22 de julio, 2011

[Per00] Pérez-Cruz, F.: *Máquina de Vectores Soporte Adaptativa y Compacta*. Tesis doctoral, Universidad Politécnica de Madrid, Madrid, España, 2000

[PNR+99] Pérez-Cruz, F., Navia-Vázquez, Á., Rojo-Álvarez, J. L. y Artés-Rodríguez, A.: *A New Training Algorithm for Support Vector Machines*, en: *Proceedings of the Fifth Bayona Workshop on Emerging Technologies in Telecommunications*, pp. 116-120. Baiona, España, 1999

[PS00] Pena-Reyes, C. A. y Sipper, M.: *Evolutionary Computation in Medicine: An Overview*. Journal of Artificial Intelligence in Medicine 19(1), 1-23, 2000

[Pur05] Purcell, G. P.: *What Makes a Good Clinical Decision Support System*. BMJ, 330 740/1, 2005

[RR04] Rioseco, M. L. y Riquelme, R. O.: *Neumonía Neumocócica Bacterémica en 45 Adultos Inmunocompetentes Hospitalizados. Cuadro Clínico y Factores Pronósticos*. Rev Méd Chile, 132:588-594, 2004

[SC04] Shawe-Taylor, J. y Cristianini, N.: *Kernel Methods for Pattern Analysis*. Cambridge University Press, Nueva York, Nueva York, Estados Unidos, 2004

[SGG+01] Sim, I., Gorman, P., Greenes, R., Haynes, R., Kaplan, B., Lehmann, H. Y Tang. P.: *Clinical Decision Support Systems for the Practice of Evidence-Based Medicine*. Journal of the American Medical Informatics Association, 8(6), 527/4, 2001

[Smo96] Smola, A. J.: *Regression Estimation with Support Vector Learning Machine*. Tesis de máster, Technische Universität München, Munich, Alemania, 1996

[SS01] Schölkopf, B. y Smola, A. J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts, Estados Unidos, 2001

[SS02] Schölkopf, B. y Smola, A. J.: *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

[SS04] Smola, A. J. y Schölkopf, B.: *A Tutorial on Support Vector Regression*. Statistics and Computing, 14(3), pp. 199-222, 2004

[SSB+81] Shortliffe, E. H., Scott, A. C. Bischoff, M. B., Campbell, A. B., Melle, W. va. y Jacobs, C. D.: *Oncocin: An Expert System for Oncology Protocol Management*, en: *Seventh International Joint Conference on Artificial Intelligence*, Vancouver, 1981

[SV99] Suykens, J. A. K. y Vandewalle, J.: *Least Squares Support Vector Machine Classifiers*. Neural Processing Letters, 9 (3), 293-300, 1999

[SVD02] Suykens, J., Van Gestel, T., De Brabanter, J., De Moor, B., y Vandewalle, J.: *Least Squares Support Vector Machines.* World Scientific Publishing, Singapore, 2002

[SWS+00] Schölkopf, B., Williamson, R. C., Smola, A., Shawe-Taylor, J. y Platt, J.: *Support Vector Method for Novelty Detection*. Advances in Neural Information Processing Systems, vol. 12, 2000

[TD04] Tax, D. M. J., Duin, R.: *Support Vector Data Description*. Machine Learning, pp.45-66, 2004

[TK03] Theodoridis, S. y Koutroubas, K.: *Pattern Recognition*. Londres: Academia, Press, 2003

[TZC+09] Tang, Y., Zhang, Y. Q., Chawla, N. V. y Krasser, S.: *SVMs Modeling for Highly Imbalanced Classification*. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics, 39(1), pp. 281-288, 2009

[Vap82] Vapnik, V. N.: *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, Nueva York, Estados Unidos 1982

[Vap95] Vapnik, V. N.: *The Nature of Statistical Learning Theory*. Springer-Verlag, Nueva York, Nueva York, Estados Unidos, 1995

[Vap98] Vapnik, V. N.: *Statistical Learning Theory*. Wiley-Interscience, Nueva York, Nueva York, Estados Unidos, 1998

[Vap99] Vapnik, V. N.: *An Overview of Statistical Learning Theory*. IEEE Transac- tions on Neural Networks, 10(5), pp. 988-999, 1999

[VKK08] Varonen, H., Kortteisto, T., y Kaila, M.: *What May Help or Hinder the Implementation of Computerized Decision Support Systems (CDSSS): a Focus Group Study with Physicians*. Family Practice Advance Access, 2008

[VL63] Vapnik, V. N. y Lerner, A.: *Pattern Recognition Using Generalized Portrait Method*. Automation and Remote Control, 24(6), pp. 774-780, 1963

[WC03] Wu, G. y Chang, E. Y.: *Class-Boundary Alignment for Imbalanced Dataset Learning*, en: *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, pp. 49-56. Washington, Distrito de Columbia, Estados Unidos, 2003